

**Instituto Tecnológico de Costa Rica**  
**Escuela de Ingeniería en Electrónica**



**Detección automática de bandas en imágenes de geles de electroforesis  
por medio de la optimización de una función objetivo**

**Informe de Proyecto de Graduación para optar por el título de Ingeniero  
en Electrónica con el grado académico de Licenciatura**

**David Soto Vásquez**

**Cartago, Junio de 2010**

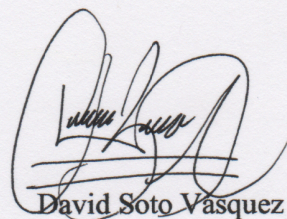


Declaro que el presente Proyecto de Graduación ha sido realizado enteramente por mi persona, utilizando y aplicando literatura referente al tema e introduciendo conocimientos propios.

En los casos en que he utilizado bibliografía, he procedido a indicar las fuentes mediante las respectivas citas bibliográficas.

En consecuencia, asumo la responsabilidad total por el trabajo de graduación realizado y por el contenido del correspondiente informe final.

Cartago, 23 Junio 2010



David Soto Vázquez

Céd: 1-1334-0923



**Instituto Tecnológico de Costa Rica**


**Escuela de Ingeniería Electrónica**

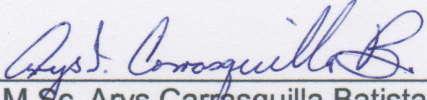
**Proyecto de Graduación**

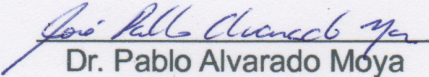
**Tribunal Evaluador**

Proyecto de Graduación defendido ante el presente tribunal evaluador como requisito para optar por el título de Ingeniería en Electrónica con el grado académico de Licenciatura, del Instituto Tecnológico de Costa Rica.

**Miembros del tribunal**

  
M.Sc. Eduardo Interiano Salguero  
Profesor Lector

  
M.Sc. Arys Carrasquilla Batista  
Profesora Lectora

  
Dr. Pablo Alvarado Moya  
Profesor Asesor

Los miembros de este tribunal dan fe de que el presente trabajo de graduación ha sido aprobado y cumple con las normas establecidas por la Escuela de Ingeniería Electrónica.

Cartago, 23 de junio del 2010



# Resumen

El presente proyecto contribuye al diseño e implementación de algoritmos orientados al análisis de imágenes de geles de electroforesis, con el objetivo de facilitar la interpretación de la información que estas imágenes contienen. Esta información se encuentra representada por la ubicación de las bandas dentro de cada carril. Sin embargo aspectos como la limitación de la percepción visual humana, el bajo contraste de las imágenes y la falta de un criterio unificado de lo considerado como banda dificultan su ubicación manual.

Por este motivo el presente trabajo tiene como objetivo principal la ubicación automática de las bandas presentes en las imágenes de geles de electroforesis. La solución propuesta se basa en la optimización de una función objetivo, utilizando como optimizador el método *Downhill Simplex* y algoritmos genéticos para la generación de puntos multidimensionales a optimizar. La función objetivo utilizada es la función de error medio cuadrático entre la imagen del carril y una sumatoria de funciones gaussianas que modelan la distribución de los fragmentos de proteínas o ADN/ARN a lo largo de cada carril, de tal forma que al encontrar los parámetros, entiéndose amplitud, media y varianza de las funciones gaussianas que mejor ajusten la función a la imagen del carril, se encuentra la ubicación central de las bandas y su respectivo nivel de intensidad.



# Abstract

This project contributes to the design and implementation of algorithms oriented to the analysis of electrophoresis gel images. These are designed for easier interpretation of the information contained in these images. This information is represented by the location of the bands in each lane. However, some aspects like limitations of the human visual perception, low contrast of the gel images and the nonexistence of one unique criterion of what should be considered as band, make the manual location of the bands harder.

This project's main goal is to automatically find the location of each band in electrophoresis gel images. The proposed solution is based in the optimization of a target function using the Downhill Simplex method and the use of genetic algorithms to generate multidimensional points to be optimized. This function represents the mean quadratic error between the image of the lane and the sum of gaussian functions that model the distribution of the DNA/RNA or proteins along each lane. Finding the best set of parameters (amplitude, mean and variance) that best fit the sum to the lane image, the intensity level and the central location of each band are obtained.

---

**Keywords:** *Electrophoresis Gel, Downhill Simplex, Gaussian function, Genetic algorithms.*



*A mi amada madre...*

# Agradecimiento

Primeramente a Dios por permitirme llegar a finalizar mi carrera y por darme la fuerza para superar todos los obstáculos presentados a lo largo de ella.

A mi madre Flor Vásquez Jimenez a quién dedico el presente trabajo por apoyarme y creer siempre en mí.

Muchas gracias a mis hermanas Vanessa y Anayancy por el apoyo que me han brindado en todos aquellos momentos difíciles de mi vida y a mi novia Mary Cruz por apoyarme y alentarme a seguir adelante con la realización de este proyecto.

Al Dr. Pablo Alvarado Moya por toda la ayuda brindada desde el inicio de este proyecto, por sus invaluable consejos a lo largo de la realización del mismo, muchas gracias.



# Índice General

Índice Figuras.....	iii
Índice Tablas.....	v
Lista de símbolos y abreviaciones.....	vi
1 Introducción.....	1
1.1 Caracterización molecular de organismos.....	1
1.2 Problemática existente en el análisis de las imágenes de geles de electroforesis.....	3
1.3 Solución propuesta para la detección automática de las bandas presentes en los carriles de las imágenes de los geles de electroforesis.....	4
1.4 Objetivos y estructura de éste trabajo.....	5
2 Marco Teórico.....	6
2.1 Función Sigmoide, Función Gaussiana y Error cuadrático medio.....	6
2.1.1 Función Sigmoide.....	6
2.1.2 Función Gaussiana.....	8
2.1.3 Error cuadrático medio.....	9
2.2 Optimización de parámetros.....	10
2.2.1 Downhill Simplex.....	11
2.2.2 Gradientes conjugados.....	13
2.3 Algoritmos genéticos y Frentes de Pareto.....	15
2.3.1 Evaluación multiobjetivo: Frentes de Pareto.....	15
2.3.2 Algoritmo evolutivo multi-objetivo: PESA.....	17
2.4 Umbralización.....	18
2.5 Trabajos anteriores realizados para la detección automática de bandas.....	18
3 Detección automática de bandas.....	20
3.1 Extracción del carril y selección de información a analizar.....	22
3.2 Estimación de $\sigma$ .....	24
3.3 Extracción del fondo del carril y segmentación en ventanas.....	27
3.3.1 Extracción del fondo del carril.....	27
3.3.2 Segmentación del carril en ventanas.....	27
3.4 Optimización de la función objetivo.....	28
3.4.1 Definición de la función objetivo.....	28
3.4.2 Estrategia de optimización.....	30
3.5 Procesamiento de las bandas encontradas.....	34

3.5.1 Umbralización y organización de bandas.....	34
3.5.2 Corrección del duplicado de bandas entre ventanas.....	35
3.5.3 Corrección del duplicado de bandas en el vector de parámetros final .....	36
4 Análisis de Resultados.....	38
4.1 Estimación de $\sigma$ .....	39
4.2 Extracción del fondo según $\sigma$ .....	41
4.3 Optimización de la función objetivo.....	43
4.3.1 Cantidad de bandas en la imagen del carril igual a la cantidad de bandas utilizadas por la función objetivo.....	44
4.3.2 Cantidad de bandas en la imagen del carril mayor a la cantidad de bandas utilizadas por la función objetivo.....	45
4.3.3 Cantidad de bandas en la imagen del carril menor a la cantidad de bandas utilizadas por la función objetivo.....	46
4.4 Umbralización.....	48
4.5 Corrección del duplicado de bandas en las regiones de traslape.....	49
4.6 Fusión de bandas.....	51
4.7 Medición de la desviación promedio de las bandas estimadas.....	51
4.8 Rendimiento.....	53
4.9 Análisis de carriles completos de geles de electroforesis.....	54
5 Conclusiones y Recomendaciones.....	56
5.1 Conclusiones .....	56
5.2 Recomendaciones .....	57



# Índice Figuras

Figura 1.1: Esquema del proceso de electroforesis.....	2
Figura 1.2: Bloques principales del módulo de procesamiento de imágenes.....	3
Figura 1.3: Gel de electroforesis.....	4
Figura 1.4: Diagrama de bloques general de la solución propuesta.....	5
Figura 2.1: Función Sigmoide y su derivada.....	7
Figura 2.2: Función Gaussiana.....	8
Figura 2.3: Representación geométrica de una superficie de error.....	10
Figura 2.4: Movimientos del Simplex.....	12
Figura 2.5: Direcciones de búsqueda.....	14
Figura 2.6: Ilustración del concepto de dominio en un frente de Pareto.....	16
Figura 3.1: Diagrama de flujo del sistema propuesto.....	21
Figura 3.2: Gel de electroforesis con distorsión de efecto barril.....	23
Figura 3.3: Recorrido del carril por ventanas con traslape completo.....	24
Figura 3.4: $\sigma$ en función del número de iteración.....	26
Figura 3.5: Recorrido del carril con ventanas en cascada.....	28
Figura 3.6: Funciones. (a) Sigmoide $A(\psi)$ y (b) Sigmoide $\mu(\omega)$ .....	29
Figura 3.7: Diagrama modular de la estrategia de optimización de la función objetivo para el caso de la intensidad promedio.....	31
Figura 3.8: Frente de Pareto para el análisis de una fila promedio.....	32
Figura 3.9: División de la ventana para corregir duplicado de bandas.....	35
Figura 3.10: Duplicado de bandas en los límites de la ventana.....	36
Figura 3.11: Aproximación de una banda real mediante varias bandas estimadas.....	37
Figura 4.1: Carril sintético utilizado para evaluar la extracción del fondo.....	41
Figura 4.2: Carril sintético sin fondo resultante.....	41
Figura 4.3: Extracción del Fondo.....	42
Figura 4.4: Distribución de intensidad de la fila central del carril.....	42
Figura 4.5: Carril sintético para la evaluación de la optimización de la función objetivo.....	43
Figura 4.6: Distribuciones de intensidad considerando una cantidad de bandas igual a las de la ventana. (a)Distribución del carril, (b)Distribución estimada .....	45
Figura 4.7: Distribuciones de intensidad, mayor cantidad de bandas en la ventana que las consideradas. (a) real, (b) Estimada.....	46
Figura 4.8: Carril de referencia. (a) Imagen del carril (b)Distribución de intensidad promedio.....	47

Figura 4.9: Carril estimado. (a) Imagen del carril (b)Distribución de intensidad estimada.....	48
Figura 4.10: Carril utilizado para pruebas de umbralización.....	49
Figura 4.11: Estimación de la ubicación central de bandas sin utilizar umbralización.....	49
Figura 4.12: Estimación de la ubicación central de bandas con umbralización.....	49
Figura 4.13: Segmento de carril utilizado para la evaluación del duplicado de bandas debido al traslape .....	49
Figura 4.14: Segmentación en ventanas con duplicado de bandas en la region de traslape (a) Primera ventana (b) Segunda ventana.....	50
Figura 4.15: Carril sintético para la evaluación de la fusión de bandas.....	51
Figura 4.16: Carril estimado sin considerar la fusión de bandas.....	51
Figura 4.17: Carril estimado considerando la fusión de bandas.....	51
Figura 4.18: Carriles involucrados en la estimación promedio de las bandas.....	53
Figura 4.19: Carril extraído de un gel de electroforesis para evaluar el comportamiento del algoritmo ante aglomeración de bandas.....	54
Figura 4.20: Carriles involucrados en el proceso de optimización de la función objetivo para el caso de aglomeración de bandas.....	54
Figura 4.21: Carril extraído de un gel de electroforesis para evaluar el comportamiento del algoritmo si hay mayor cantidad de regiones libres de bandas .....	54
Figura 4.22: Carriles involucrados en el proceso de optimización de la función objetivo para el caso en el cual las regiones sin bandas son mayores que las regiones con bandas.....	55

# Índice Tablas

Tabla 4.1: Valores reales y estimados de $\sigma$ para carriles sintéticos.....	39
Tabla 4.2: Medidas estadísticas para las estimaciones de $\sigma$ de los carriles sintéticos.....	40
Tabla 4.3: Anchos originales y estimados de las bandas de los carriles sintéticos.....	40
Tabla 4.4: Valores de intensidad para las bandas en un carril sintético con fondo y sin fondo.....	41
Se observa que para el caso del carril sintético solo es eliminado del carril lo considerado como fondo, el cual también es eliminado de la amplitud máxima de las bandas.....	41
Tabla 4.5: Parámetros de PESA y Downhill simplex para la optimización.....	43
Tabla 4.6: Valores reales y estimados de los parámetros de las bandas, considerando todas las bandas presentes en la ventana.....	44
Tabla 4.7: Valores reales y estimados de intensidad y ubicación para la optimización que considera menor cantidad de bandas.....	45
Tabla 4.8: Valores reales y estimados de intensidad y ubicación para la optimización de un carril sintético que considera mayor cantidad de bandas.....	47
Tabla 4.9: Parámetros de bandas estimados para un carril real con menos bandas que las consideradas por la función objetivo.....	48
Tabla 4.10: Ubicación central de las bandas estimadas sin el algoritmo de corrección del traslape.....	50
Tabla 4.11: Ubicación central de las bandas estimadas con el algoritmo de corrección del traslape.....	50
Tabla 4.12: Ubicación de bandas real y estimada para la medición de la desviación promedio de las bandas.....	52
Tabla 4.13: Iteraciones y tiempo promedio necesarios para la detección automática de bandas en carriles sintéticos.....	53



# Lista de símbolos y abreviaciones

## Notación general

$\mathbf{Y}$  Matriz

$$\mathbf{Y} = \begin{bmatrix} \psi_0 & \omega_0 & \dots & \psi_{B_{pv}-1} & \omega_{B_{pv}-1} \\ \psi_0 & \omega_0 & \dots & \psi_{B_{pv}-1} & \omega_{B_{pv}-1} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \psi_0 & \omega_0 & \dots & \psi_{B_{pv}-1} & \omega_{B_{pv}-1} \end{bmatrix}$$

$\mathbf{P}$  Vector

$$\mathbf{P} = [\psi_0 \ \omega_0 \ \psi_1 \ \omega_1 \ \dots \ \psi_{B_{pv}-1} \ \omega_{B_{pv}-1}]^T$$

## Símbolos

$\Theta$	Vector final de parámetros de bandas
$N_b$	Número de bandas encontradas en un carril
$S_t$	Tamaño del kernel para la apertura morfológica
$\varphi$	Factor de proporción para $S_t$
$w$	Número de columnas de la ventana a analizar
$M_v$	Ventana con mayor varianza de intensidad
$B_{pv}$	Número de bandas por ventana
$A$	Intensidad en el píxel central de la banda
$\mu$	Ubicación central de la banda
$\psi$	Mapeo de $A$ a la representación sigmoideal
$\omega$	Mapeo de $\mu$ a la representación sigmoideal
$\mathbb{P}^N$	Espacio de parámetros
$\mathcal{P}_i$	Población interna
$\mathcal{P}_E$	Población externa
$F_{\mathcal{P}_i}$	Fenotipos de la población interna
$S_{\mathcal{P}_i}$	Población de individuos localmente optimizados
$\mathbb{C}$	Cromosoma de un individuo

$P_c$	Probabilidad de cruce de individuos
$P_m$	Probabilidad de mutación de individuos
$\Lambda$	Vector de aptitud
$\varepsilon$	Valor de la función ECM
$\xi$	Tolerancia del Downhill Simplex
$M_d$	Mediana
$\nabla$	Operador gradiente
$T$	Valor umbral
$R$	Región
$\mu_t$	Ubicación central de la banda fusionada
$S_c$	Carril sintético
$E_c$	Carril estimado
$ui$	Unidades de intensidad
$\Delta\mu$	Tolerancia para fusión de bandas

## Abreviaciones

ECM	Error cuadrático medio
PBV	Parámetros de bandas en la ventana
PESA	Pareto Envelope-based Selection Algorithm

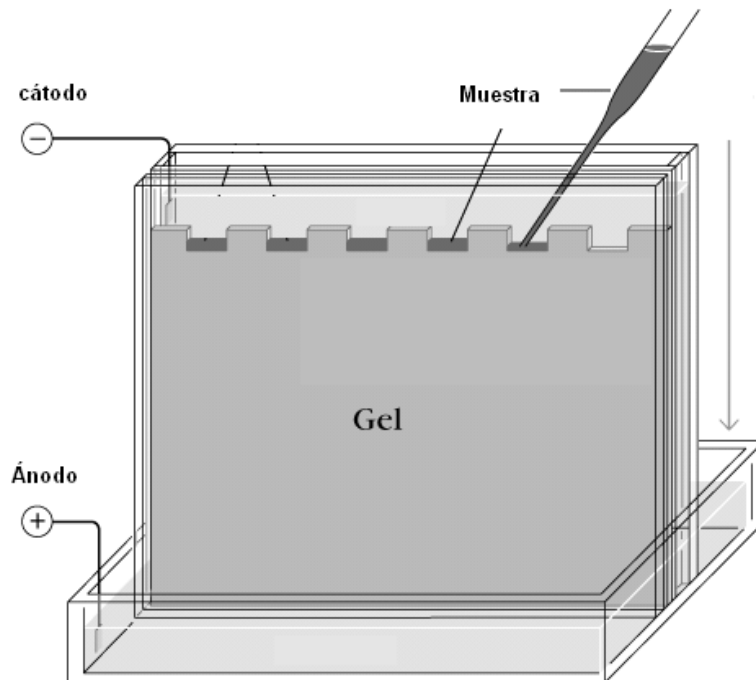
# Capítulo 1

## 1 Introducción

### 1.1 Caracterización molecular de organismos

La electroforesis en gel es una técnica ampliamente utilizada por biólogos moleculares para analizar la separación molecular de sustancias como proteínas y ácidos nucleicos, basándose en la capacidad de movilidad sobre un gel de cada una de las moléculas según su tamaño y carga eléctrica [1].

En este proceso sustancias como el ADN/ARN o proteínas son combinadas con una mezcla de enzimas de restricción con el objetivo de realizar una segmentación de las moléculas presentes. Posteriormente la muestra segmentada es inyectada en una matriz formada por un gel generalmente de poliacrilamida o de agarosa, que son dos tipos distintos de polímeros, los cuales funcionan como soporte o base para el análisis a realizar. La figura 1.1 ilustra el proceso.



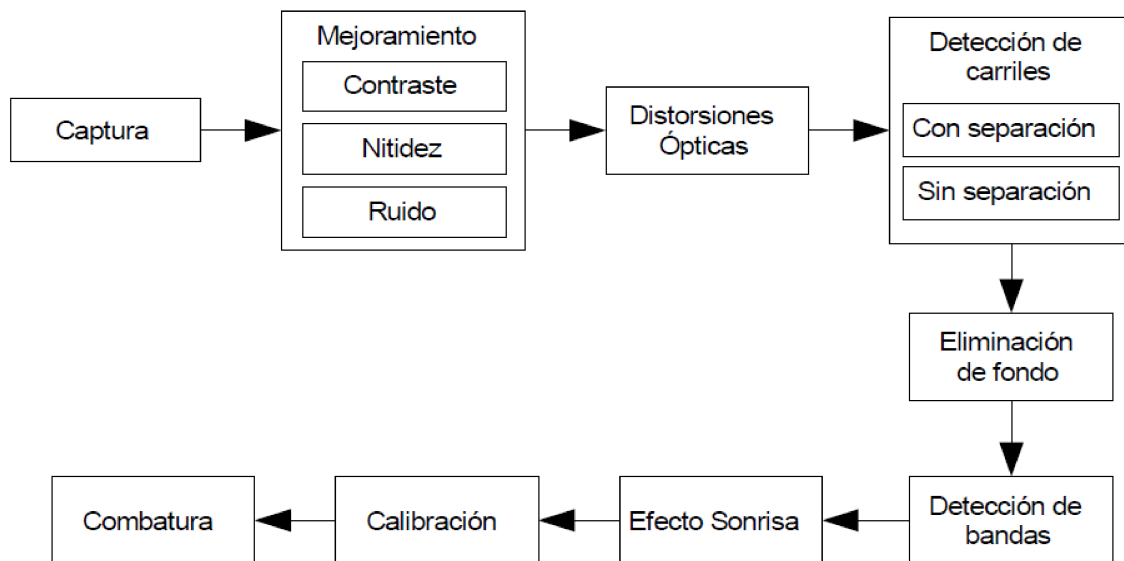
**Figura 1.1:** Esquema del proceso de electroforesis

El gel utilizado como soporte se encuentra sometido a un campo eléctrico generado de forma externa y es el causante del desplazamiento de las moléculas a lo largo del gel. Para el caso de los ácidos nucleicos como el ADN en los cuales predominan las cargas negativas, las moléculas se desplazan del cátodo de la fuente que genera el campo eléctrico al ánodo. Este desplazamiento es directamente proporcional a la cantidad de carga eléctrica de cada molécula e inversamente proporcional a su tamaño o masa [2]. Una vez realizado el proceso de electroforesis las moléculas más pequeñas y de mayor carga se encuentran a menor distancia del ánodo. Al realizar el desplazamiento sobre el gel las moléculas forman carriles hasta llegar a un punto de equilibrio en el cual la resistencia ejercida por el material no permite que la molécula continúe su movimiento, en este punto se forma una banda.

Este procedimiento permite analizar el grado de similitud entre dos muestras de ADN, lo cual posibilita su uso, por ejemplo, para pruebas de paternidad, así como para evaluar en pruebas de clonación el nivel de semejanza entre dos secuencias de ADN. Esta es una de las aplicaciones dadas al método en el entorno costarricense.

Este trabajo es parte de un proyecto de investigación que contribuye al análisis de imágenes de geles de electroforesis y que pretende facilitar la interpretación de los resultados obtenidos al realizar la electroforesis en gel, esto mediante el desarrollo de una plataforma de asistencia para laboratorios de biología molecular que permita el manejo tanto de los procesos de captura, mejora y análisis básico de las imágenes, como el manejo de la meta-información asociada a cada carril de un gel de electroforesis, permitiendo así un análisis automático [3].

La figura 1.2 ilustra los bloques principales involucrados en el análisis automático de geles de electroforesis.



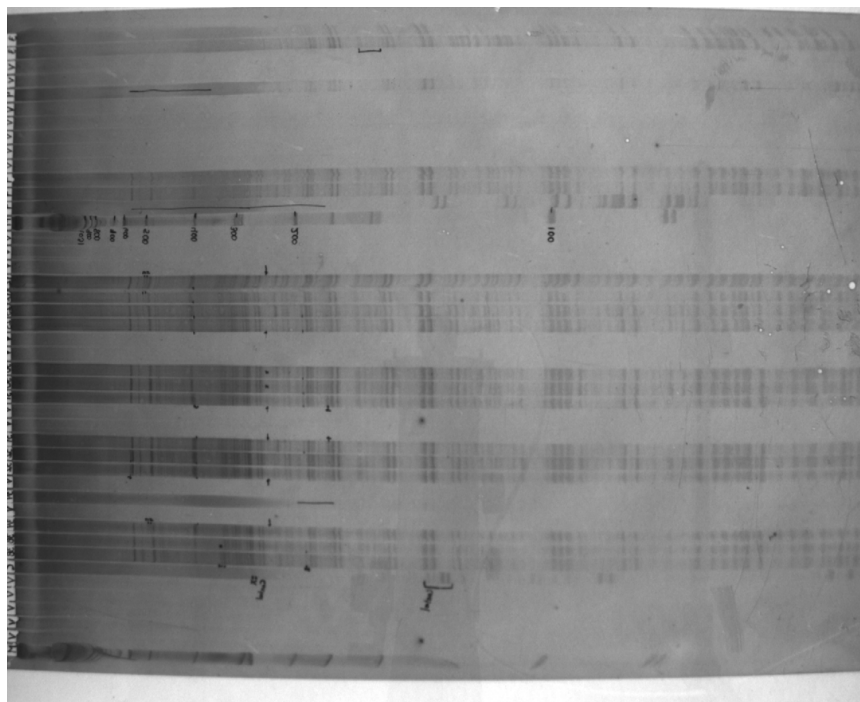
**Figura 1.2:** Bloques principales del módulo de procesamiento de imágenes [4].

El presente trabajo implementa una estrategia para realizar una ubicación automática de las bandas presentes en los carriles de los geles, por tanto se encuentra ubicado dentro del bloque *Detección de bandas*. Se hace uso además de los trabajos previos presentes en [4] y [5] ubicados en los bloques *Eliminación de fondo* y *Detección de carriles*.

## 1.2 Problemática existente en el análisis de las imágenes de geles de electroforesis.

El análisis de los resultados obtenidos después de aplicar el proceso de electroforesis se basa en la ubicación de las bandas presentes en los carriles formados por el desplazamiento de las moléculas. Esta ubicación es utilizada para realizar una comparación entre dos o más carriles, así como para calcular de forma indirecta características de la molécula que dio origen a la banda, como por ejemplo su peso molecular [6]. Además, su ubicación puede ser utilizada para corregir algunas de las distorsiones presentes en las imágenes de los geles, con estrategias como la empleada en [7] para corregir la distorsión del *efecto sonrisa*.

Sin embargo la ubicación manual de estas bandas en las imágenes de geles de electroforesis puede introducir error en el método debido a que las imágenes de los geles presentan distorsiones así como problemas de contraste tal y como se ilustra en la figura 1.3, donde las líneas horizontales con menor intensidad son los carriles y las líneas verticales con menor intensidad dentro de cada carril son las bandas.



**Figura 1.3:** Gel de electroforesis.

Otra problemática al realizar la búsqueda de las bandas presentes en un carril es la carencia de un criterio unificado de qué puede ser considerado como una banda, ya que tal y como se puede observar en la figura 1.3 ciertas regiones dentro de los carriles presentan mayores niveles de intensidad que el resto del carril, no obstante no pueden ser consideradas como bandas, ya que no presentan una distribución de intensidad característica del perfil de una banda. De manera similar otras regiones presentan aglomeración de bandas, que debido a limitaciones en la percepción visual humana pueden ser interpretadas de forma errónea como una sola banda.

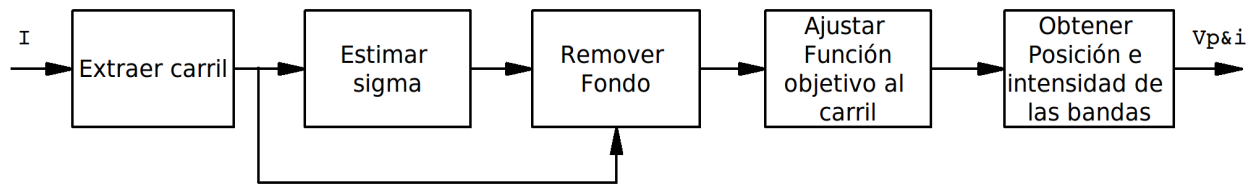
De esta forma es posible plantear la problemática mediante la siguiente pregunta: *¿Como puede realizarse una detección automática de las bandas presentes en las imágenes de los geles de electroforesis?*

### **1.3 Solución propuesta para la detección automática de las bandas presentes en los carriles de las imágenes de los geles de electroforesis.**

En el presente trabajo se expone una estrategia para realizar la ubicación automática de las bandas. Esta se basa en encontrar una función matemática que modele la distribución de moléculas o bandas que describe el carril analizado. Esta función es optimizada utilizando el algoritmo *Downhill Simplex* y *algoritmos genéticos* de tal forma que se obtenga el menor error medio cuadrático entre la



superficie que describe la función y la información contenida en la imagen del carril. La figura 1.4 ilustra un diagrama de bloques general del sistema propuesto.



**Figura 1.4:** Diagrama de bloques general de la solución propuesta

Para la solución se proponen algoritmos que han sido implementados haciendo uso del lenguaje de programación C++, así como de módulos existentes en la biblioteca de software para visión por computador y procesamiento digital de imágenes LTI-Lib. De igual forma se buscará compatibilidad en cuanto a estructura de programación con esta biblioteca.

## 1.4 Objetivos y estructura de éste trabajo

El principal objetivo del presente trabajo es ubicar de forma automática las bandas presentes en los geles de electroforesis obtenidos durante la caracterización molecular de organismos. Para esto se establecerá una función de error entre la imagen del carril que contiene las bandas a ubicar y una sumatoria de funciones gaussianas, donde cada función modela la distribución de intensidad del perfil de una banda. Esta función de error deberá ser optimizada de tal forma que se encuentre el conjunto de parámetros; amplitud, media (ubicación central de cada banda) y varianza, que minimizan la función de error. Además mediante umbralización se eliminan aquellas bandas falsas encontradas para así finalmente generar un vector multidimensional con los parámetros que describen las bandas presentes en el carril analizado.

En el capítulo 2 se detallan todos aquellos conceptos en los que se basa la solución presentada en este trabajo, conformando así el marco teórico de la solución. En el capítulo 3 se expone el diseño conceptual de la solución propuesta y en el capítulo 4 se realiza un análisis de los resultados obtenidos. Por último en el capítulo 5 se incluyen las conclusiones obtenidas mediante el análisis de resultados y se exponen algunas de las recomendaciones para trabajos posteriores .

# Capítulo 2

## 2 Marco Teórico

En el presente capítulo se detallan los fundamentos teóricos de la solución propuesta. Inicialmente se definen de forma general las funciones matemáticas utilizadas, seguidamente se realiza una explicación del concepto de optimización de parámetros y dos estrategias de optimización. Por otra parte se presentan los fundamentos de los frentes de Pareto, así como de los algoritmos genéticos y por último se finaliza con el concepto de umbralización y una reseña de los trabajos previos realizados para solucionar la problemática de la ubicación automática de las bandas.

### 2.1 Función Sigmoide, Función Gaussiana y Error cuadrático medio.

#### 2.1.1 Función Sigmoide

La función sigmoide en su forma parametrizada es definida como:

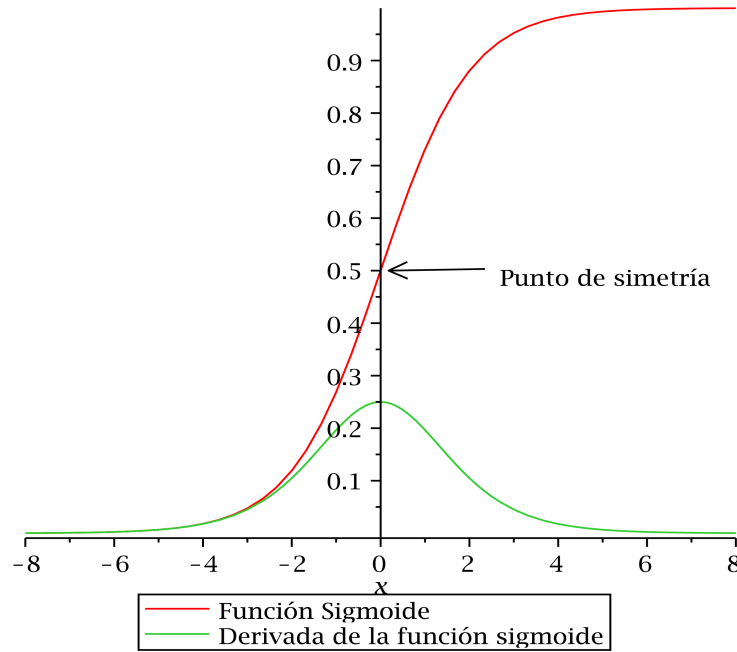
$$f(x, \alpha, \beta, \gamma) = \frac{\alpha}{(1 + \exp(-\beta x))} + \gamma \quad (2.1)$$

También es llamada función logística y se caracteriza por ser una función real y diferenciable. Se encuentra delimitada en su codominio por sus parámetros  $\alpha$  y  $\gamma$ . Su dominio está definido para todo el conjunto de los números reales.

Además su derivada siempre es positiva [8]. Se define para el caso  $\alpha=\beta=1$  y  $\gamma=0$ , como:

$$f'(x, \alpha, \beta, \gamma) = (1 - f(x, \alpha, \beta, \gamma))f(x, \alpha, \beta, \gamma) \quad (2.2)$$

Por otra parte la función sigmoide satisface la condición  $f(x, \alpha, \beta, \gamma) + f(-x, \alpha, \beta, \gamma) = 1$  y como se ilustra en la figura 2.1 tiende rápidamente a su límite inferior conforme su argumento disminuye, de igual forma sucede con su límite superior, cuando su argumento crece [9]. Por otra parte posee un comportamiento lineal para argumentos cercanos a su punto de simetría.



**Figura 2.1:** Función Sigmoide y su derivada ( $\alpha=\beta=1, \gamma=0$ ).

Es posible establecer el rango o codominio de la función así como su tasa de crecimiento y punto de simetría modificando sus parámetros, cuyas funciones son las siguientes:

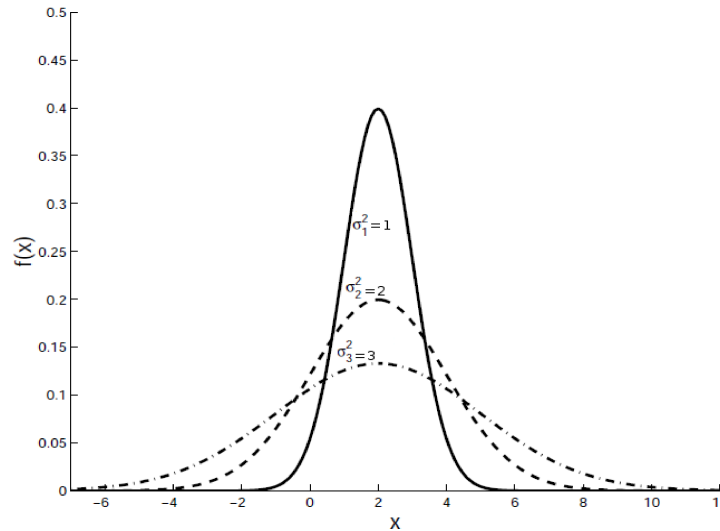
- $\alpha$  : Este parámetro establece el rango dinámico de la función [8].
- $\beta$  : Ajusta la pendiente de la transición entre el mínimo y el máximo de la función.
- $\gamma$  : Permite ajustar el valor del punto de simetría, al realizar un desplazamiento de la función y establecer el valor mínimo y máximo del sigmoide.

## 2.1.2 Función Gaussiana

Se define una variable aleatoria como aquella función que asigna o relaciona un número perteneciente a un espacio muestral a cada resultado de un experimento [10]; puede ser discreta o continua. Ésta se encuentra distribuida normalmente si tiene una función de densidad de probabilidad descrita por:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.3)$$

Esta función es denominada función o distribución gaussiana, con valor medio  $\mu$  y varianza  $\sigma^2$ . La función gaussiana es simétrica respecto a su parámetro  $\mu$  en el cual alcanza su valor máximo y describe una curva en forma de campana con una dispersión según su varianza [11]. Un 99.74% de esta dispersión se encuentra comprendida en el intervalo  $[\mu - 3\sigma, \mu + 3\sigma]$  tal y como se ilustra en la figura 2.2.



**Figura 2.2:** Función Gaussiana [11].

La función gaussiana está completamente descrita a través de sus dos parámetros  $\mu$  y  $\sigma^2$  los cuales son sus momentos de primer y segundo orden respectivamente. Para el caso de variables aleatorias discretas, es posible encontrar su valor utilizando las siguientes ecuaciones:

$$\mu = \frac{\sum_{i=0}^n x_i}{n} \quad (2.4)$$

$$\sigma^2 = \frac{\sum_{i=1}^n x_i^2 - n\mu^2}{n} \quad (2.5)$$

En [12] se modela la distribución de fragmentos que conforman una banda en la electroforesis capilar en gel, mediante una función gaussiana, caracterizada por una varianza formada principalmente por cuatro factores:

$$\sigma_T^2 = \sigma_{iny}^2 + \sigma_{det}^2 + \sigma_{\Delta T}^2 + \sigma_{dif}^2 \quad (2.6)$$

donde  $\sigma_{iny}$  representa la dispersión de la banda causada por la inyección de la muestra,  $\sigma_{det}$  representa el efecto causado por el instrumento con el cual se realiza la detección de las bandas, es decir el aporte del fenómeno de detección a la dispersión,  $\sigma_{\Delta T}$  corresponde el aporte debido al gradiente térmico dentro del capilar y por último  $\sigma_{dif}$  considera la difusión en el gel.

Es posible generalizar los resultados obtenidos en [12] para el caso de la electroforesis en gel, sin embargo, para este caso el aporte debido a la detección no es considerado. De igual forma sucede con el gradiente térmico ya que es posible despreciar su aporte para simplificar el análisis, dando como resultado que la varianza total de la banda se encuentra definida por la suma de las contribuciones individuales de la inyección de la muestra y la difusión debido al gel [13]:

$$\sigma_T^2 = \sigma_{iny}^2 + \sigma_{dif}^2 \quad (2.7)$$

En el presente trabajo se adopta el criterio utilizado por Luckey *et al.* en [12] al modelar la distribución espacial de los fragmentos de ADN que conforman una banda mediante una función gaussiana. En este caso la distribución de intensidad que representa el perfil de una banda dentro de la imagen del carril, tiene un máximo según el nivel de intensidad de la banda y un valor medio que representa la ubicación o píxel en la cual se encuentra el centro de la banda. Por último se considera que todas las bandas presentes en un carril poseen un mismo valor de varianza.

### 2.1.3 Error cuadrático medio

El error cuadrático medio entre dos señales discretas  $I_1$  e  $I_2$  se obtiene mediante:

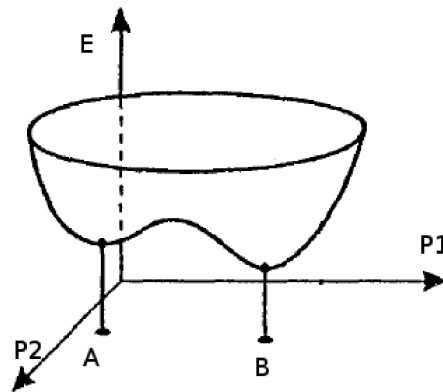
$$ECM(I_1, I_2) = \frac{1}{N} \sum_1^N (I_1 - I_2)^2 \quad (2.8)$$

Donde N es la cantidad de muestras que componen a ambas señales. El error cuadrático medio en el área de procesamiento de señales es utilizado como un criterio de evaluación de la fidelidad y de la calidad de una señal y es utilizado como criterio base en algoritmos de optimización de señales ya que posee propiedades como simetría y convexidad, además de ser una función diferenciable. Algunas de las ventajas de esta función de error son detalladas en [14].

## 2.2 Optimización de parámetros

La optimización de una función que depende de uno o varios parámetros se basa en la búsqueda de los valores para estos de tal forma que maximicen o minimicen la función, la cual es llamada función objetivo [15]. Para el caso en el cual la función objetivo es una función de error, la optimización consiste en encontrar el conjunto de parámetros que minimicen la función, es decir que permitan alcanzar ya sea un mínimo local o un mínimo global según los requisitos de la aplicación.

Es posible explicar el proceso de optimización de parámetros desde un punto de vista geométrico [16], considerando la función objetivo como una superficie de error en un espacio N-dimensional, donde cada una de estas dimensiones corresponde a un parámetro de la función tal y como se ilustra en la figura 2.3. La optimización consiste en encontrar el punto multidimensional que represente las coordenadas del mínimo global (B) o local (A) sobre la superficie de error.



**Figura 2.3:** Representación geométrica de una superficie de error [16].

Las estrategias de optimización de parámetros se pueden dividir principalmente en dos clases: la primera de ellas se basa solamente en la información obtenida mediante evaluaciones de la función objetivo y la segunda utiliza además la información brindada por la derivada de la función para el caso unidimensional o su correspondiente gradiente para el caso multidimensional.

En lo restante de esta sección se detallan dos estrategias de optimización de parámetros utilizadas durante la realización de este trabajo. La primera de ellas es llamada *Downhill Simplex* la cual se basa únicamente en evaluaciones de la función objetivo y la segunda; llamada *Gradientes Conjugados*, que además de la función objetivo utiliza la información de su gradiente.



## 2.2.1 Downhill Simplex

El método de optimización *Downhill Simplex* fue introducido por Nelder y Mead en [17]. Este método, a diferencia de otros algoritmos de optimización, no utiliza en su interior una estrategia de optimización en una dimensión, ya que el concepto utilizado trabaja simultáneamente con todas las dimensiones involucradas debido al enfoque geométrico del algoritmo, basado en el movimiento de un *simplex* a través de la superficie de error.

Un simplex es una figura geométrica compuesta por  $N+1$  vértices, con  $N$  la dimensión o cantidad de parámetros del problema de optimización a tratar. Cada vértice del simplex corresponde a un punto de  $N$  dimensiones. Para el caso de dos y tres dimensiones, los simplex utilizados por el algoritmo son un triángulo y un tetraedro respectivamente. El método recibe un simplex inicial como punto de partida para realizar la optimización. En general este simplex debe ser no degenerado, es decir debe encerrar un volumen finito [15].

Estableciendo un punto inicial  $P_0$  para uno de los vértices del simplex, es posible obtener los restantes vértices mediante la siguiente ecuación:

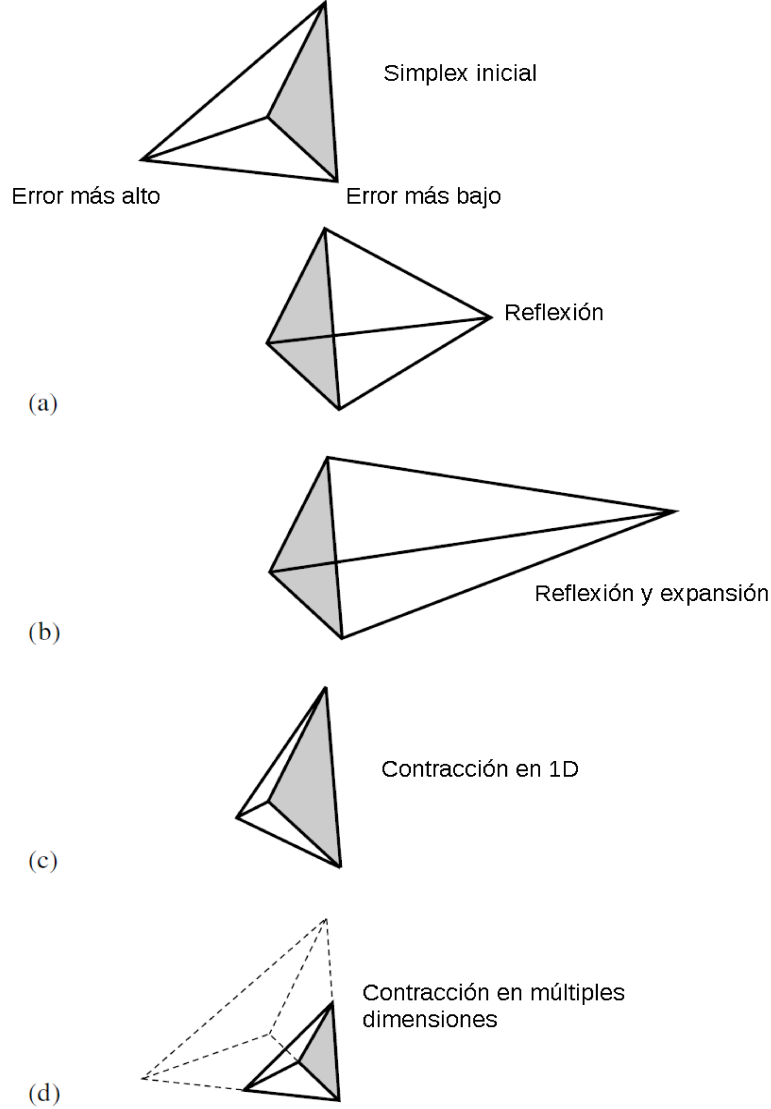
$$P_i = P_0 + \lambda e_i \quad (2.9)$$

donde  $e_i$  son  $N$  vectores unitarios y  $\lambda$  es una constante que depende de la escala del problema de optimización.

El método se basa en la comparación de los valores de la función objetivo para cada uno de los vértices del simplex, sustituyendo en cada iteración el vértice que posee el valor de error más alto. En cada paso, el simplex puede realizar cuatro diferentes movimientos sobre la superficie de error: reflexión, expansión, contracción en 1D y contracción en todas la dimensiones hacia el vértice con menor error, estos cuatro movimientos son ilustrados en la figura 2.4.

El proceso inicia ubicando el vértice que posee mayor error  $P_{may}$ , y calculando el centroide  $\bar{P}$  del simplex con los  $N$  vértices restantes mediante la siguiente ecuación:

$$\bar{P} = \frac{\sum_{i=1, i \neq may}^N P_i}{N} \quad (2.10)$$



**Figura 2.4:** Movimientos del Simplex, (a) Reflexión, (b) Expansión, (c) Contracción en 1D y (d) Contracción en múltiples dimensiones [15].

Los movimientos de reflexión, expansión y contracción en 1D, se realizan utilizando la ecuación:

$$\zeta = \bar{P} + \alpha(\bar{P} - P_{may}) \quad (2.11)$$

donde  $\zeta$  es el punto multidimensional resultante y  $\alpha$  es una constante que determina el tipo de movimiento a realizar según su magnitud y su signo. El valor de  $\zeta$  en la función de error es representado por  $E_\zeta$ .

En cada iteración el algoritmo inicialmente realiza una reflexión del punto con mayor error, pasando a través del centroide, estableciendo  $\alpha=-1$  en (2.11). Si  $E_\zeta$  es menor que el actual mínimo dentro del simplex, se realiza una expansión ( $\alpha=-2$ ) dando como resultado un nuevo punto  $\zeta$ , si

$E_{\zeta} < E_{\varsigma}$  el vértice con mayor error es sustituido por  $\zeta$ , de lo contrario  $\varsigma$  es utilizado.

Por otra parte si  $E_{min} < E_{\varsigma} < E_{max}$ , con  $E_{min}$  y  $E_{max}$  los actuales valores mínimo y máximo dentro del simplex respectivamente,  $P_{may}$  es sustituido por  $\varsigma$ . De lo contrario se busca un nuevo punto  $\zeta$  realizando una contracción en una dimensión ( $\alpha=0.5$ ) hacia el actual vértice con menor error, si  $E_{\varsigma} < E_{\zeta}$  se realiza la respectiva sustitución de lo contrario la contracción se considera fallida y una contracción en todas las dimensiones hacia el vértice con menor error es realizada, terminando así la actual iteración.

Esta serie de pasos dan como resultado el desplazamiento total del simplex sobre la superficie de error hacia el mínimo de la función; sin embargo, este mínimo puede ser local o global. En el presente trabajo el algoritmo es reiniciado varias veces utilizando diferentes simplex para mejorar la búsqueda del mínimo global. Además se adopta el criterio de parada utilizado en [15], basado en una cantidad máxima de iteraciones permitidas o una tolerancia  $\xi$  mínima de decrecimiento para la distancia existente entre los valores de función máximo y mínimo dentro del simplex, de tal forma que la optimización se continua realizando de forma iterativa mientras se cumpla la siguiente inecuación:

$$\xi < \frac{2|E_{max} - E_{min}|}{|E_{max} + E_{min} + \epsilon|} \quad (2.12)$$

donde  $\epsilon$  es el valor más pequeño que sumado a 1 en la representación de punto flotante, produce un número diferente de 1.

## 2.2.2 Gradientes conjugados

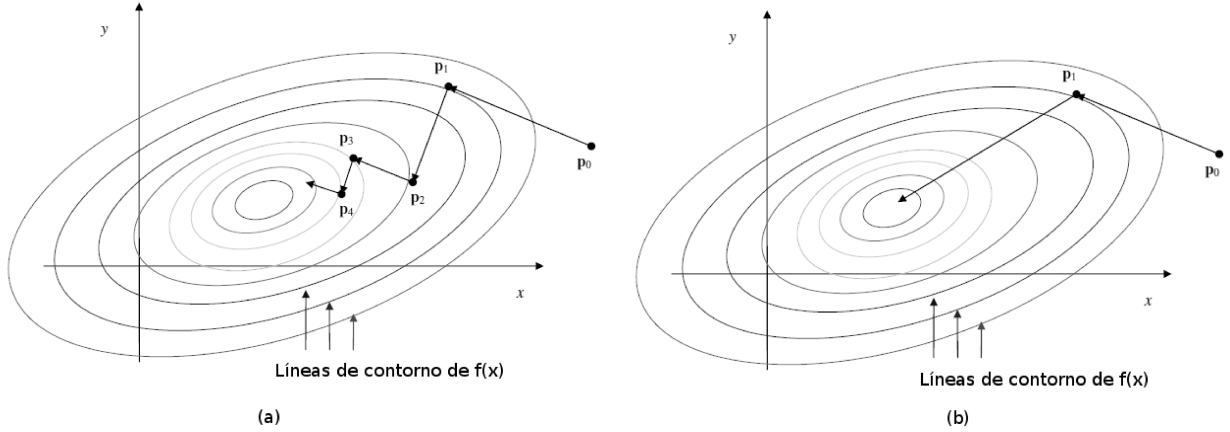
Esta estrategia de optimización consiste en un proceso iterativo, el cual se basa en la información brindada por la función de error y su gradiente, así como en la utilización de un algoritmo de optimización en una dimensión o lineal, de tal forma que el uso del mismo en cada una de las dimensiones involucradas en una optimización multidimensional permite encontrar el mínimo de una superficie de error establecida en un espacio N-dimensional [15].

La optimización en una dimensión o lineal consiste en ubicar la dirección  $\mathbf{d}^{\tau}$  en la cual la minimización se debe realizar y en seleccionar cuánto se debe desplazar un punto inicial  $\mathbf{p}^{\tau}$  tal que el mínimo en esa dirección de búsqueda sea alcanzado, es posible representar lo anterior mediante la siguiente ecuación:

$$\mathbf{p}^{\tau+1} = \mathbf{p}^{\tau} + \lambda^{\tau} \mathbf{d}^{\tau} \quad (2.13)$$

donde  $\mathbf{d}$  es la dirección de búsqueda,  $\lambda$  es la cantidad de desplazamiento,  $\tau$  la actual iteración y  $\mathbf{p}^{\tau+1}$  el mínimo alcanzado en la actual dirección.

Gradientes conjugados utiliza un conjunto de direcciones de búsqueda llamadas direcciones conjugadas para realizar la optimización. Estas direcciones son seleccionadas de tal forma que la optimización realizada en la dirección  $\mathbf{d}^T$  no sea alterada por las optimizaciones en las restantes direcciones [16], evitando así oscilaciones en los valles cercanos al mínimo de la función de error o iteraciones no necesarias [18], tal y como se ilustra en la figura 2.5 para el caso de dos dimensiones.



**Figura 2.5:** Direcciones de búsqueda (a)Dirección contraria al gradiente, (b)Direcciones conjugadas [18].

Es posible obtener un conjunto de direcciones conjugadas aproximando una función objetivo  $E$  mediante los dos primeros términos de su *serie de Taylor*:

$$E(\mathbf{x}) = E_0 + \mathbf{b}^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{H} \mathbf{x} \quad (2.14)$$

donde  $\mathbf{b}$  es constante y  $\mathbf{H}$  es la matriz *Hessiana* de la función. Obteniendo inicialmente el gradiente de esta función:

$$\nabla E = \mathbf{b} + \mathbf{H}\mathbf{x} \quad (2.15)$$

de tal forma al realizar un desplazamiento sobre la superficie de error en alguna dirección  $\mathbf{u}$ , su gradiente varia:

$$\delta(\nabla E) = \mathbf{H} \delta(\mathbf{x}) \quad (2.16)$$

Si este desplazamiento se realiza hasta el mínimo en la dirección  $\mathbf{u}$ , y se desea minimizar el error a partir de ese punto en una dirección  $\mathbf{v}$ , tal que  $\mathbf{v}$  sea una dirección conjugada de  $\mathbf{u}$ , se debe cumplir que la optimización a lo largo de  $\mathbf{v}$  no altere la componente paralela del gradiente de la previa búsqueda, esto es:

$$\mathbf{u} \cdot \delta(\nabla E) = \mathbf{u} \cdot \mathbf{H} \cdot \mathbf{v} = 0 \quad (2.17)$$

En [16] se demuestra que es posible cumplir con esta condición seleccionando la primera dirección de búsqueda como el gradiente negativo en el punto inicial ( $d_1 = -g_1$ ) y las restantes direcciones mediante la ecuación:

$$d_{j+1} = -g_{j+1} + \beta_j d_j \quad (2.18)$$

Existen dos variantes para el cálculo de  $\beta$  la primera de ellas según el algoritmo propuesto por *Fletcher-Reeves*:

$$\beta_j = \frac{g_{j+1}^T g_{j+1}}{g_j^T g_j} \quad (2.19)$$

la cual fue modificada posteriormente por *Polak-Ribiere* dando origen a la segunda forma:

$$\beta_j = \frac{g_{j+1}^T (g_{j+1} - g_j)}{g_j^T g_j} \quad (2.20)$$

La utilización de la información de la derivada para el cálculo de las direcciones conjugadas permite que para el caso de superficies de error cuadráticas con N dimensiones o parámetros el algoritmo de gradientes conjugados realice la optimización en N iteraciones. Sin embargo, debido a la dependencia de la información presente en la derivada de la función la efectividad del algoritmo se ve reducida en superficies de error con puntos de silla, ya que el algoritmo considera aquellas regiones con gradiente igual a cero como mínimos de la función.

## 2.3 Algoritmos genéticos y Frentes de Pareto

### 2.3.1 Evaluación multiobjetivo: Frentes de Pareto

Aptitud se entiende como la cualidad que hace que un objeto sea apto o idóneo para cierto fin. En el presente trabajo se busca el conjunto de parámetros para las funciones gaussianas con mayor aptitud para modelar la distribución de bandas en un carril.

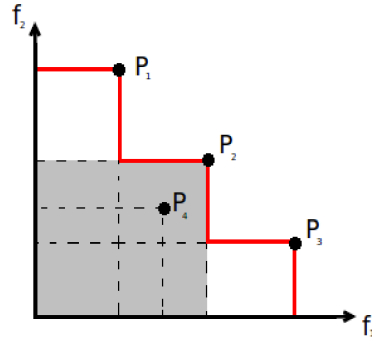
Everingham *et al.*[19] define la función de aptitud total  $F$  para un algoritmo  $A$  con una parametrización  $u$  sobre un conjunto de datos  $\mathcal{D}$  como:

$$F(A_u, \mathcal{D}) = \Phi(f_1(A_u, \mathcal{D}), \dots, f_n(A_u, \mathcal{D})) \quad (2.21)$$

donde las funciones de aptitud  $f_i(A_u, \mathcal{D})$  son definidas de tal forma que incrementen monótonicamente con algún aspecto o parámetro del algoritmo evaluado. El conjunto de funciones de aptitud evaluadas forman un espacio multidimensional de aptitudes, en el cual cada punto sobre ese espacio representa el rendimiento del algoritmo evaluado con un conjunto de parámetros.



En general la función  $\Phi$  es desconocida, pero incrementa monotónicamente con el aumento de los valores de todas las funciones de aptitud. Esto permite asegurar que un punto en el espacio de aptitud pueda ser considerado con más aptitud que otros puntos siempre que estos posean valores más pequeños en todas las dimensiones; en este caso se dice que estos puntos son *dominados* [20]. En la figura 2.6, por ejemplo, el punto  $P_4$  es *dominado* por el punto  $P_2$ , de igual forma para cualquier punto ubicado en el área sombreada. Por otra parte los puntos  $P_1$ ,  $P_2$  y  $P_3$  son puntos *no dominados* y conforman el frente de Pareto.



**Figura 2.6:** Ilustración del concepto de dominio en un frente de Pareto

Matemáticamente este concepto es expresado mediante la siguiente ecuación:

$$\hat{p} = \{ \langle \mathbf{u} \in \mathbb{P}_A, \mathbf{f}(\mathbf{A}_u, \mathcal{D}) \rangle \mid \neg \exists \mathbf{v} \in \mathbb{P}_A : \mathbf{f}(\mathbf{A}_v, \mathcal{D}) > \mathbf{f}(\mathbf{A}_u, \mathcal{D}) \} \quad (2.22)$$

donde  $\hat{p}$  es el frente de Pareto,  $\mathbf{f}$  es el vector de funciones de aptitud  $[f_1, \dots, f_n]^T$  y  $\mathbb{P}_A$  es el espacio de parámetros del algoritmo  $A$ . La relación parcial de ordenamiento  $>$  en  $\mathbf{f}$  describe la propiedad de dominancia presente en los frentes de Pareto y se define como:

$$\mathbf{f}(\mathbf{A}_v, \mathcal{D}) > \mathbf{f}(\mathbf{A}_u, \mathcal{D}) \Leftrightarrow \forall i : f_i(\mathbf{A}_v, \mathcal{D}) \geq f_i(\mathbf{A}_u, \mathcal{D}) \wedge \exists i : f_i(\mathbf{A}_v, \mathcal{D}) > f_i(\mathbf{A}_u, \mathcal{D}) \quad (2.23)$$

Considerando que el espacio de parámetros  $\mathbb{P}_A$  por lo general contiene una cantidad infinita de elementos o parametrizaciones, es necesaria una estrategia para seleccionar un grupo de muestras *representativas* de dicho espacio tal que el frente de Pareto obtenido sea una representación confiable del frente de Pareto que considera todos los elementos del espacio  $\mathbb{P}_A$ . Para esto en la solución propuesta presentada se utiliza la variante realizada en [20] al algoritmo evolutivo multi-objetivo PESA (*Pareto Envelope-based Selection Algorithm* [21]), empleando así un enfoque genético que permite suprimir el uso de parametrizaciones inútiles, enfocándose solamente en aquellas regiones pertenecientes a  $\mathbb{P}_A$  que proporcionan resultados prometedores.

### 2.3.2 Algoritmo evolutivo multi-objetivo: PESA

PESA es un algoritmo evolutivo utilizado para obtener los elementos *no dominados* que conforman los frentes de Pareto. Considera a cada posible parametrización como un *individuo*, representado dentro del algoritmo mediante un *fenotipo*, que corresponde a su representación parametrizada, y mediante un *cromosoma* que es su representación binaria.

La búsqueda de los elementos no dominados se basa en los principios de mutación y cruce de los individuos que presentan mayor aptitud. La mutación busca mejorar un individuo a través de cambios aleatorios, invirtiendo algún bit de su cromosoma si un número aleatorio entre 0 y 1 extraído de una distribución uniforme es más pequeño que la tasa de mutación  $P_m$ , basándose así en la búsqueda de candidatos dentro del vecindario de alguno de los actuales individuos no dominados. Por otra parte el cruce de individuos toma dos elementos no-dominados y los combina para generar un tercero. En esta combinación cada bit del elemento resultante es heredado de sus padres con la misma probabilidad [20].

A un conjunto de individuos se le llama *población*. La estrategia utilizada por el algoritmo mantiene dos poblaciones: la población externa  $p_E$  y la población interna  $p_I$ . La población externa son todos los individuos que se encuentran en el actual frente de Pareto. La población interna, es generalmente más pequeña que la población externa y es el conjunto de candidatos que serán evaluados para determinar si son eventualmente incluidos en la población externa.

El algoritmo mantiene un control de la cantidad de individuos que conforman la población externa, esto con el objetivo de evitar que exceda su tamaño máximo. Para esto se utiliza una estrategia de selección de individuos a eliminar. PESA con el objetivo de intentar mantener un frente de Pareto igualmente distribuido realiza un seguimiento del nivel de densidad de individuos en diferentes regiones del espacio de aptitud y elimina aquellos que se encuentren en las regiones con mayor densidad. Para realizar esta medición de densidad, el espacio de aptitud es particionado en hiper-cajas, manteniendo para cada una un factor correspondiente a su densidad. Sin embargo, en la presente solución no se utiliza el concepto de hiper-cajas y en su lugar se utiliza un kernel estimador de densidad [20]. Para el caso de la selección de los individuos a cruzar o mutar, la estrategia es opuesta, ya que se da prioridad a aquellos individuos que se encuentren en las regiones con menor densidad.

La probabilidad de cruce  $P_c$  se define como la fracción de nuevos individuos que son generados mediante el cruce de dos fenotipos de la población externa. Por lo tanto  $1 - P_c$  establece la probabilidad con la que se realiza la mutación de individuos.

Alvarado [20] propone una variación al algoritmo original, permitiendo a la tasa de mutación  $P_m$  decrecer asintóticamente a partir de un valor inicial  $P_{m_{inicial}}$  hacia la tasa de mutación final deseada  $P_{m_{final}}$ , definiendo la tasa de mutación total como:

$$P_m = (P_{m_{inicial}} - P_{m_{final}})e^{-i/\tau} + P_{m_{final}} \quad (2.24)$$

donde  $i$  es el número de iteración y  $\tau$  el factor de decrecimiento de la tasa de mutación. De esta forma es posible obtener una mayor tasa de mutación en las iteraciones iniciales, dando como resultado un muestreo aleatorio más fuerte del espacio de parámetros cuando los valores de los parámetros son relativamente inestables. Al aumentar la cantidad de iteraciones es deseable realizar un aprovechamiento de la información contenida en el actual frente en lugar de buscar nuevos puntos, es decir una menor tasa de mutación, esto también es logrado con la anterior modificación.

## 2.4 Umbralización

La umbralización es una técnica utilizada para separar una imagen en regiones o contornos correspondientes a objetos de interés [22]. Esta es basada en la intensidad de dichas regiones, eliminando o reduciendo a un valor mínimo de intensidad toda aquella región de la imagen que posea un nivel de intensidad por debajo de un valor establecido como umbral. Las regiones con un nivel de intensidad superior al umbral pueden mantener su nivel original de intensidad o ser ajustadas a un valor deseado.

Si  $T$  es el valor de umbral utilizado, es posible expresar el proceso de umbralización como:

$$g(y, x) = \begin{cases} f(y, x) & \text{si } f(y, x) \geq T \\ 0 & \text{en caso contrario} \end{cases}$$

## 2.5 Trabajos anteriores realizados para la detección automática de bandas

Se encontraron tres tendencias principales para resolver el problema de la ubicación automática de las bandas presentes en los geles de electroforesis:

En [23] la problemática se resuelve según el análisis molecular a realizar. La estrategia del método se basa en realizar un análisis de carriles orientados verticalmente, recorriendo con una plantilla en forma de “sonrisa” (forma de banda a ubicar) todo el carril, posicionando la plantilla en cada una de las filas de la imagen del carril, y calculando para cada fila en la cual la plantilla es ubicada el valor medio de los píxeles que la plantilla abarca, creando con cada uno de los valores medios obtenidos después de recorrer todo el carril una proyección. Se toma como criterio que la ubicación de las bandas se encuentra en los valles o mínimos locales de la proyección obtenida. Para facilitar la ubicación de los mínimos, inicialmente se ecualiza el histograma de intensidades, posteriormente se realiza una segmentación morfológica y por último la proyección se filtra con un filtro pasa bajas para eliminar las restantes irregularidades. El algoritmo finaliza con la ubicación de los mínimos locales de la proyección resultante. Sin embargo, esta estrategia parte de la premisa de que todas las bandas presentes en el carril han sufrido distorsión y presentan forma de “sonrisa” lo cual no siempre es cierto. Además depende en gran parte de la plantilla utilizada y no es capaz de detectar efectivamente las bandas en sectores con aglomeración de bandas ya que los identifica como una única banda.

Bajla et al.[24] destaca la problemática existente con el diseño de una estrategia de detección de bandas totalmente automática y propone una técnica basada en dos etapas, que consideran la información del carril en sus dos dimensiones y la interacción con el usuario. En la primera etapa la imagen del carril es regularizada en intensidad utilizando un filtro GDD (*Geometry Driven Diffusion*), seguidamente con los carriles verticalmente orientados se aplica a la imagen un detector lineal de los límites horizontales de las bandas, el cual es un acumulador de diferencias de intensidad entre las filas, definido como:

$$D_i = \sum_j |I_{i+1,j} - I_{ij}| \quad (2.25)$$

donde,  $I_{ij}$  es el nivel de intensidad del píxel ubicado en la fila  $i$  y columna  $j$ . Posteriormente se realiza una ubicación de los máximos locales en  $D_i$ . El rectángulo creado entre cada par de máximos locales es considerado una banda. De esta forma finaliza la primera etapa y el usuario puede añadir o eliminar bandas a las actuales encontradas. Finalmente en la segunda etapa se utiliza el gradiente de la imagen en el vecindario de los píxeles que se detectaron como los bordes de las bandas con el objetivo de mejorar la ubicación de los límites. Para esto se forma un rectángulo que involucra el límite o borde dado por el gradiente, parte de la región considerada como fondo y parte de la región considerada como banda, obteniendo el indicador del límite final de la banda como la diferencia absoluta entre la media de la región del fondo y la media de la región de la banda. Debido a la regularización de intensidad realizada en la primera etapa del algoritmo, esta estrategia no permite realizar la ubicación de bandas que se encuentran en regiones del carril con bajo contraste, ya que bandas con bajos niveles de intensidad serán consideradas como fondo a pesar de poseer el perfil gaussiano de intensidad de una banda.

Por último en [25] y [26] se proponen dos estrategias basadas en la deconvolución mediante la estimación de parámetros utilizando el método estadístico de *máxima verosimilitud* (*maximum likelihood*). La primera de ellas para un carril verticalmente orientado, crea un carril con el promedio de intensidad de cada fila del carril original y posteriormente busca ajustarlo a uno de los perfiles de un genotipo previamente almacenado en una base de datos. No obstante, no siempre se cuenta con una base de datos que incluya el perfil del genotipo a analizar, lo cual limita la utilidad del método. La segunda busca ajustar la representación en 1D  $y(s)$  de la secuencia de ADN con una sumatoria de funciones deltas de Dirac:

$$x(s) = A_0 + \sum_{j=1}^p A_j \delta(s - \tau_j) + e \quad (2.26)$$

considerando que la señal  $y(s)$  se puede obtener mediante:

$$y(s) = x(s) * w(s) \quad (2.27)$$

donde  $w(s)$  es la respuesta al impulso, la estrategia se enfoca en encontrar las posiciones centrales de las bandas  $\tau_j$  y su respectiva amplitud  $A_j$  [4]. Sin embargo, se asume que no existen bandas en los extremos del carril y que existe ruido blanco  $e$  normalmente distribuido con baja varianza en todo el carril, lo cual no es cierto en las imágenes de los geles.

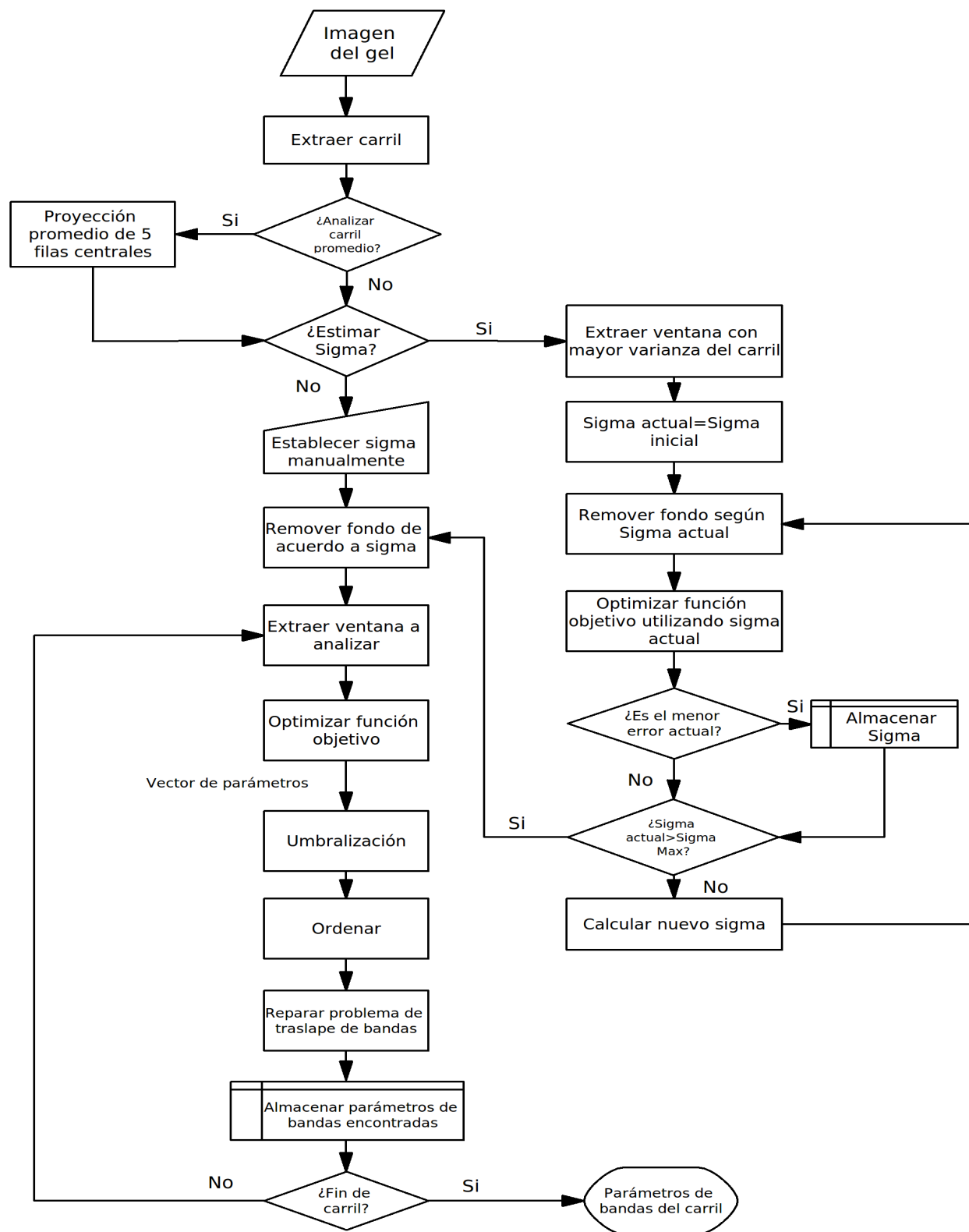
# Capítulo 3

## 3 Detección automática de bandas

En la figura 3.1 se muestra un diagrama de flujo de la solución propuesta para la detección de bandas en imágenes de geles de electroforesis. Ésta se basa en el análisis individual de cada carril horizontal presente en la imagen de geles, específicamente en la distribución de intensidad del carril. El método es capaz de realizar un análisis basado en la proyección promedio de las cinco filas centrales del carril sobre un vector, o de analizar en forma paralela la distribución de intensidad de una cantidad de filas establecidas por el usuario. Estas filas son seleccionadas de la fila central hacia los extremos del carril.

La estrategia se fundamenta en la forma gaussiana típica descrita por el perfil de intensidad de las bandas, tal que, la distribución de intensidad presente a lo largo de un carril se puede modelar como una sumatoria de funciones gaussianas, cuyos parámetros describen a cada una de las bandas, esto eliminando previamente lo considerado como fondo de la imagen del carril. En la solución propuesta se adopta como criterio de diseño que todas las bandas presentes en el carril tienen una misma varianza o  $\sigma$ , la cual es un parámetro de entrada para el algoritmo. Sin embargo esta varianza puede ser estimada por medios algorítmicos y en el presente trabajo se propone un método para ello. Los restantes parámetros, amplitud y valor medio, son estimados mediante el uso de un método de optimización de parámetros en conjunto con el algoritmo PESA y los frentes de Pareto, utilizando como función objetivo una función de error medio cuadrático entre la distribución de intensidad del carril y la sumatoria de funciones gaussianas.





**Figura 3.1:** Diagrama de flujo del sistema propuesto

Debido a que el problema de optimización involucra 2 dimensiones por cada posible banda, se propone realizar un análisis del carril por ventanas para disminuir la cantidad de iteraciones necesarias para converger al mínimo de la función objetivo, ya que estas como se explica más adelante están en función de la cantidad total de dimensiones del problema de optimización. En el análisis realizado dos ventanas consecutivas se traslapan una distancia de  $8\sigma$ , para evitar pérdida de bandas. El grupo de posibles bandas o funciones gaussianas ubicadas por el método es sometido a segmentación mediante umbralización. Por último, las bandas resultantes son ordenadas según su ubicación y una estrategia para solucionar duplicados de bandas es aplicada. Como resultado del proceso se obtiene un vector  $\Theta$  de  $N$  dimensiones ( $N=2N_b$  con  $N_b$  el número de bandas encontradas en todo el carril) que contiene los parámetros de intensidad y ubicación central de cada posible banda localizada en el carril.

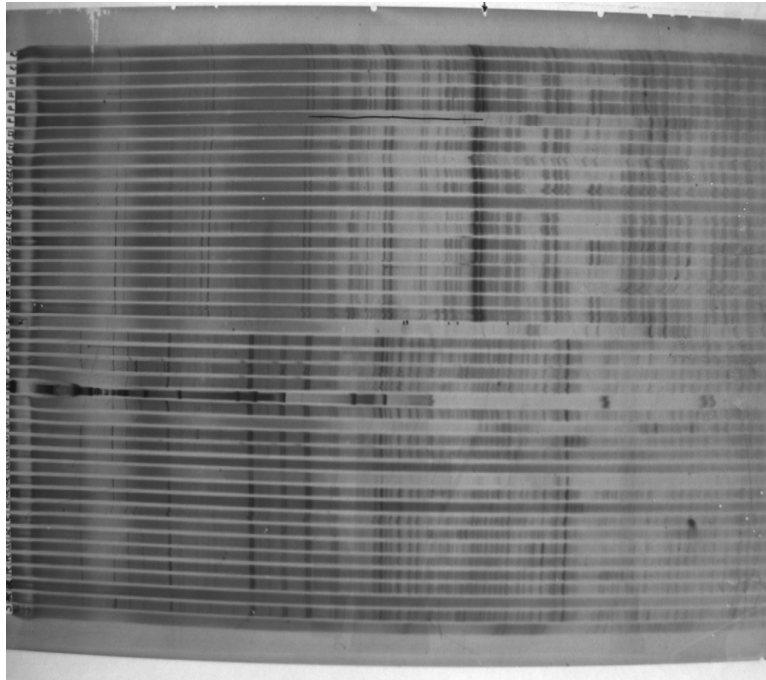
En lo restante de esta sección los detalles involucrados en el desarrollo de la solución propuesta son explicados, siendo posible dividir el diseño en las siguientes etapas:

- Extracción del carril y selección de información a analizar.
- Estimación de  $\sigma$ .
- Extracción del fondo del carril y segmentación en ventanas.
- Optimización de la función objetivo.
- Procesamiento de las bandas encontradas.

### **3.1 Extracción del carril y selección de información a analizar**

Las imágenes de los geles de electroforesis presentan diversos tipos de distorsiones, entre estas: la distorsión geométrica y el efecto “sonrisa”, que se derivan tanto de la naturaleza del método como de la mecánica involucrada en el proceso de captura de las imágenes.

Estas distorsiones dificultan la extracción de los carriles presentes en las imágenes de los geles, principalmente la distorsión geométrica, la cual es un tipo de distorsión óptica producida por el arreglo de lentes en el objetivo de la cámara, dando como resultado un aspecto redondeado a la imagen, afectando así su geometría. La figura 3.2 muestra una imagen de un gel de electroforesis con distorsión geométrica conocida como efecto barril.



**Figura 3.2:** Gel de electroforesis con distorsión de efecto barril

En el presente trabajo se utiliza la solución propuesta en [5] para corregir inicialmente la distorsión geométrica presente en las imágenes de los geles y posteriormente realizar la extracción de los carriles.

Obtenida la imagen del carril de dimensiones  $n \times m$ , con  $m$  filas y  $n$  columnas, se realiza un análisis de la distribución de intensidad. Este análisis puede ser realizado de dos formas diferentes. La primera de ellas considerando de forma paralela la distribución de intensidad de  $p$  filas a partir de la fila central de la imagen del carril, siendo  $p$  un parámetro establecido manualmente. Este tipo de análisis da como resultado  $p$  vectores  $V_k$  con  $k \in \{0, 1, 2, \dots, p-1\}$ , cada uno de dimensión

$N$ . Por otra parte, cada banda encontrada  $b_u$  es descrita por los parámetros de amplitud  $A_u$  y media  $\mu_u$  de su respectiva función gaussiana. Así para el caso de la banda  $b_0$ , sus parámetros  $A_0$  y  $\mu_0$  en cada una de las filas en consideración se encontrarán respectivamente en la primera y segunda dimensión de cada vector  $V_k$ . Para este caso el algoritmo da como resultado que los parámetros de las bandas ubicadas dentro del carril corresponden a la mediana  $M_d$  de cada una de las  $N$  dimensiones de los  $p$  vectores.

Lo anterior puede ser expresado en forma matricial, considerando una matriz  $Y$  de dimensión  $p \times N$  ( $N=2N_b$ ) formada por cada uno de los  $p$  vectores. El vector resultante  $V_f$  que contiene los parámetros que describen a las bandas presentes en el carril se obtiene de la mediana de cada una de las columnas de  $Y$ .

$$Y = \begin{bmatrix} A_0 & \mu_0 & \dots & A_{N_b} & \mu_{N_b} \\ A_0 & \mu_0 & \dots & A_{N_b} & \mu_{N_b} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ A_0 & \mu_0 & \dots & A_{N_b} & \mu_{N_b} \end{bmatrix} \quad (3.1)$$

$$V_f = [M_d(Y_{C_0}) \quad M_d(Y_{C_1}) \quad \dots \quad M_d(Y_{C_{N-2}}) \quad M_d(Y_{C_{N-1}})]^T \quad (3.2)$$

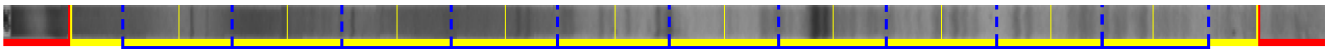
La segunda forma de análisis consiste en realizar una proyección sobre la fila central de la imagen del carril, utilizando ésta como vector de intensidades de referencia para la función objetivo a optimizar. Esta proyección consiste en asignar a la posición  $i$ , con  $i \in \{0, 1, 2, \dots, n-1\}$  del vector central el promedio de los valores de intensidad del elemento  $i$  de las cinco filas centrales de la imagen del carril. De esta forma se disminuye el tiempo de procesamiento del primer método y se considera la información presente en las filas vecinas de la fila central. Lo anterior se expresa matemáticamente para cada posición  $i$  como:

$$D_i = \frac{1}{5} \sum_{j=j_c-2}^{j=j_c+2} D_{j,i} \quad (3.3)$$

donde  $D_{j,i}$  corresponde al valor de la fila  $j$  y columna  $i$  de la imagen del carril y  $j_c$  la posición de la fila central.

## 3.2 Estimación de $\sigma$

La estrategia utilizada para la estimación de la desviación estándar o  $\sigma$  realiza inicialmente una división del carril en ventanas de tamaño finito, esto con el objetivo de encontrar la ventana en que se ubica la mayor cantidad de bandas. La figura 3.3 ilustra la división realizada. Se utiliza una división en ventanas completamente traslapadas para evitar la pérdida de información entre dos ventanas consecutivas. Para este análisis no se considera el 5% inicial y final de la imagen del carril, esto debido a que en la mayoría de las imágenes procesadas estos sectores se encuentran con niveles de intensidad similares a los de las bandas, sin embargo no corresponden a bandas, sino a efectos no deseados durante el proceso de electroforesis. Estos sectores son señalados por las ventanas con contorno de color rojo en la figura 3.3.



**Figura 3.3:** Recorrido del carril por ventanas con traslape completo. Las líneas de color azul indican las ventanas con traslape consideradas sobre las ventanas amarillas.

En cada una de las ventanas se utiliza como indicador de presencia de bandas la varianza de la intensidad presente en la fila central de la ventana. La varianza para una ventana de dimensiones  $w \times m$  se calcula mediante:

$$\sigma_v^2 = \frac{\sum_{i=0}^{w-1} D_i^2 - w \mu_v^2}{w} \quad (3.4)$$

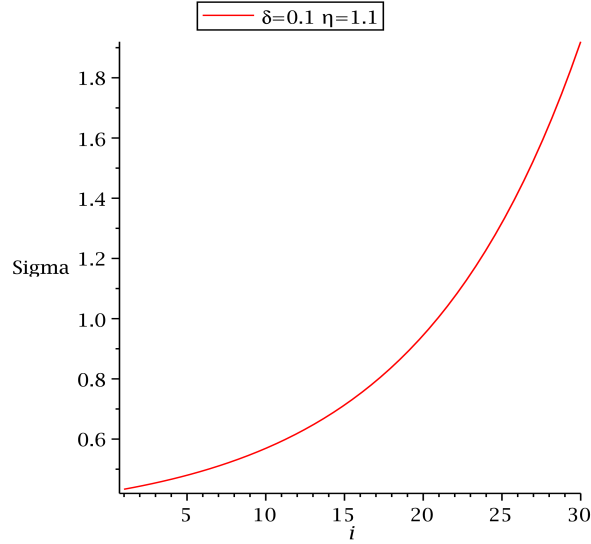
donde  $v$  indica el índice de la ventana en análisis. Se asume que la ventana con mayor variación en su nivel de intensidad contiene la mayor cantidad de bandas.

La ventana con mayor varianza es sometida a un proceso iterativo en búsqueda de la varianza que mejor se ajusta al perfil de intensidad de las bandas que contiene. Esto se realiza por medio de optimizaciones de la función objetivo. El procedimiento utilizado para realizar la optimización es explicado en detalle en la sección 3.4. Este proceso recibe como parámetro de entrada el  $\sigma$  de las bandas a encontrar y da como resultado además de los parámetros de cada banda encontrada, un factor de *aptitud* que indica el nivel de ajuste entre los parámetros de las bandas estimadas y la distribución real de intensidad de la fila central de la ventana. En cada iteración se elimina lo considerado como fondo en la ventana, esto de acuerdo al  $\sigma$  de la iteración. De esta forma el presente método de estimación de  $\sigma$  se basa en buscar en un intervalo establecido de posibles varianzas, el valor que dé como resultado el mayor factor de aptitud.

El recorrido del intervalo  $[\sigma_{inicial}, \sigma_{final}]$  de las posibles desviaciones estándar de las bandas presentes en la ventana se realiza siguiendo la siguiente ecuación:

$$\sigma_i = \sigma_{inicial} + \delta \eta^{i-1} \quad \forall i \neq 0 \quad (3.5)$$

donde  $i$  corresponde al número de iteración y las variables  $\delta$  y  $\eta$  definen la tasa de variación de la desviación estándar entre cada iteración. La figura 3.4 ilustra el comportamiento de esta ecuación para los valores seleccionados en la presente solución. Al asignar la desviación estándar en cada iteración mediante (3.5) se permite que una mayor cantidad de valores sean evaluados para la primera mitad del intervalo de posibles valores. En la presente solución se selecciona el intervalo  $[1/3, 10]$  esto como consecuencia de un análisis manual de las varianzas típicas presentes en las bandas de las imágenes de geles de electroforesis utilizadas.



**Figura 3.4:**  $\sigma$  en función del número de iteración

Para eliminar el fondo de la ventana se utiliza el algoritmo propuesto en [4] y explicado en la sección 3.3. Este se basa en la apertura morfológica, para ello se utiliza un elemento estructurador  $S$  de tamaño definido por:

$$S_t = \sigma_i \varphi \quad (3.6)$$

donde  $\sigma_i$  es la desviación estándar utilizada durante la iteración  $i$ , multiplicado por un factor  $\varphi$  que es seleccionado de tal forma que la ventana resultante contenga sólo las bandas más un ruido atenuado. En el método implementado se selecciona  $\varphi=7$  considerando que las bandas poseen un perfil de intensidad gaussiano y por ende el tamaño de la banda es aproximadamente  $6\sigma$ .

Así, el Algoritmo 1 resume la propuesta para la estimación de la desviación estándar de las bandas presentes en el carril.

---

**Algoritmo 1 para la estimación de  $\sigma$**

---

*Entrada:* Imagen del carril  $I(x, y)$  de tamaño  $n \times m$

*Salida:*  $\sigma$  estimada,  $\sigma_e$

1. Remover 5% inicial y final de  $I(x, y)$
2. Dividir la imagen en ventanas de tamaño  $w \times m$
3. **Para cada** ventana
4. Calcular la varianza de intensidad en su fila central

$$\sigma_v^2 = \frac{\sum_{i=0}^{w-1} D_i^2 - w \mu_v^2}{w}$$

5. **Si**  $\sigma_v$  es la mayor actual
  6. Guardar índice de la ventana  $M_v = v$
-

7. **fin si**
8. **fin para**
9.  $\sigma_i = \sigma_{\text{inicial}}$
10. **mientras**  $\sigma_i \leq \sigma_{\text{final}}$
11. *Remover fondo de la ventana*  $M_v$   
 $\text{kernel}_{\text{tamaño}} = \sigma_i \varphi$
12. *Optimizar función objetivo sobre*  $M_v$  *con*  $\sigma_i$
13. **Si** *es el mayor factor de aptitud obtenido*
14. *Guardar  $\sigma$  actual*  
 $\sigma_e = \sigma_i$
15. **fin si**
16. *Calcular*  $\sigma_{i+1}$   
 $\sigma_{i+1} = \sigma_{\text{inicial}} + \delta \eta^i$
17. **fin mientras**

### 3.3 Extracción del fondo del carril y segmentación en ventanas

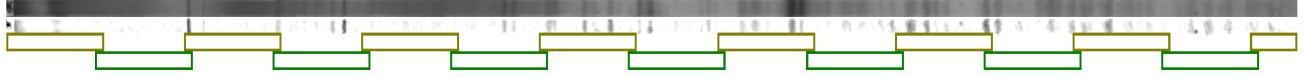
#### 3.3.1 Extracción del fondo del carril

A partir del valor de  $\sigma$  estimado o establecido de forma manual, se elimina el fondo del carril a analizar. Para esto se utilizaron los módulos desarrollados en [4] que permiten remover lo considerado como *no-banda* mediante la aplicación de diferentes filtros a la imagen. Principalmente se realiza una apertura morfológica [27] a la imagen del carril, que consiste en un filtro de mínimos seguido por un filtro de máximos. Además se aplica un filtro promediador para “suavizar” la imagen resultante, un filtro de mediana para atenuar el ruido y por último se realiza una mejora del contraste del carril. Para la apertura morfológica se selecciona un elemento estructurador de tamaño  $12\sigma$ , lo cual permite considerar aquellas regiones en las que se encuentran dos o más bandas traslapadas.

#### 3.3.2 Segmentación del carril en ventanas

El carril con el fondo previamente eliminado es segmentado en  $h$  ventanas de tamaño finito  $w \times m$ , con  $w$  un parámetro del método. Esto permite disminuir la cantidad de iteraciones necesarias por el algoritmo de optimización de parámetros Downhill Simplex para realizar el análisis del carril completo, ya que la cantidad de iteraciones totales  $i_t$  necesarias para converger al mínimo de la función objetivo está en función de la cantidad de dimensiones del problema de optimización y de la superficie de error descrita por la función objetivo, y para el peor de los casos esta función es exponencial.

Así dividiendo el análisis total del carril en ventanas y limitando la cantidad máxima de bandas por ventana ( $B_{pv}$ ) se logra acelerar la ubicación de las bandas presentes a lo largo todo del carril. Esta segmentación se ilustra en la figura 3.5, donde el traslape existente entre dos ventanas contiguas es de  $8\sigma$ , distancia suficiente para abarcar la distribución total de una banda.



**Figura 3.5:** Recorrido del carril con ventanas en cascada

El recorrido en cascada permite considerar en el análisis las bandas que se encuentran en los límites de la ventana y evita el análisis innecesario de regiones que ya han sido analizadas por el algoritmo optimizador, permitiendo disminuir el tiempo de procesamiento, en comparación con un recorrido como el de la figura 3.3. En la sección 3.5 se implementa un algoritmo para corregir la duplicación de bandas en las regiones de traslape de ventanas.

## 3.4 Optimización de la función objetivo

En esta sección se define la función objetivo a optimizar, así como la estrategia de optimización y los criterios de diseño adoptados para ubicar los parámetros de amplitud y media de las funciones gaussianas que modelan la distribución de intensidad de las bandas presentes en la ventana analizada. Lo restante de esta sección se divide en las siguientes subsecciones:

- Definición de la función objetivo
- Estrategia de optimización

### 3.4.1 Definición de la función objetivo

Inicialmente se define la función objetivo a optimizar para una ventana de tamaño  $w \times m$  como la función de error cuadrático medio ( $ECM$ ) entre una de las filas del segmento de carril encerrado en la ventana y una sumatoria de funciones gaussianas que representan a cada posible banda dentro del segmento analizado. Lo anterior es expresado matemáticamente como:

$$ECM = \frac{1}{w} \sum_{i=0}^{w-1} \left[ D(i) - \sum_{u=0}^{B_{pv}-1} A_u e^{\frac{-(i-\mu_u)^2}{2\sigma^2}} \right]^2 \quad (3.7)$$

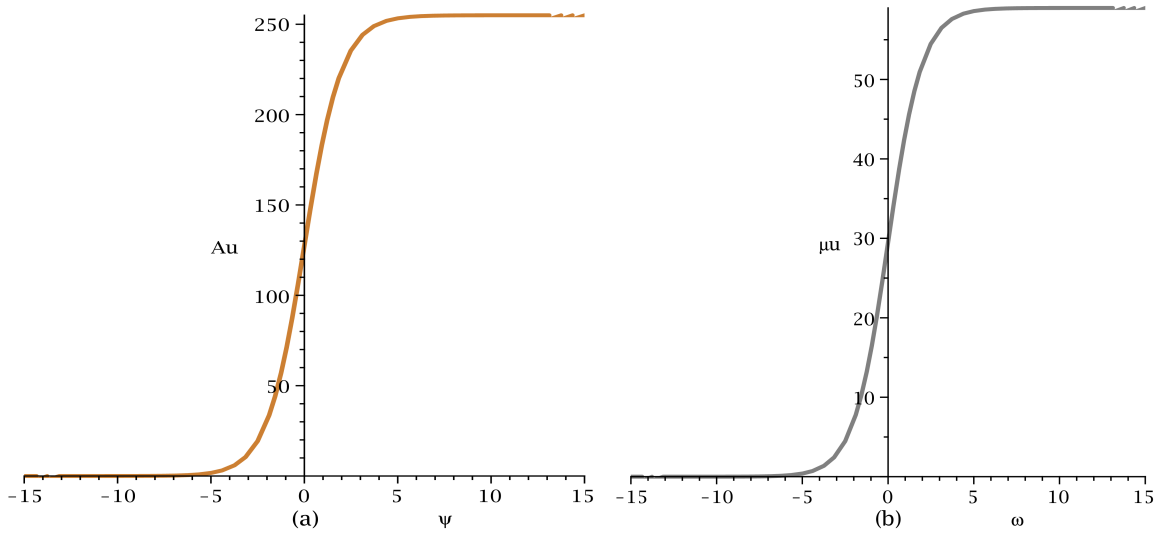
donde  $D(i)$  es el valor  $i$  de la fila analizada y el parámetro  $A_u$  el nivel de intensidad de la banda  $u$  en su píxel central  $\mu_u$ , siendo estos dos últimos los parámetros optimizados por el algoritmo, ya que  $\sigma$  es determinada en las etapas previas.



Conociendo el rango de posibles valores para los parámetros a optimizar, se realiza un cambio de variable, tal que la función de error excluya aquellos valores que no se encuentran en los rangos permitidos, esto es posible sustituyendo cada uno de estos parámetros por una función sigmoide. En este trabajo las imágenes de geles analizadas se encuentran en una escala de gris de 0 a 255, con 255 el color blanco. De esta forma el rango de posibles valores para  $A_u$  es el intervalo  $[0,255]$  y para  $\mu_u$  el intervalo  $[0,w-1]$ . La figura 3.6 ilustra el comportamiento de los cambios de variable realizados, matemáticamente definidos como:

$$A_u(\psi) = \frac{255}{1 + e^{-\psi_u}} \quad (3.8)$$

$$\mu_u(\omega) = \frac{w-1}{1 + e^{-\omega_u}} \quad (3.9)$$



**Figura 3.6:** Funciones. (a) Sigmoide  $A(\psi)$  y (b) Sigmoide  $\mu(\omega)$  con  $w=60$ .

Finalmente reescribiendo la función objetivo utilizada por el algoritmo optimizador se obtiene:

$$ECM = \frac{1}{w} \sum_{i=0}^{w-1} \left[ (255 - D(i)) - \sum_{u=0}^{B_{pv}-1} \left( \frac{255}{1 + e^{-\psi_u}} \right) e^{\frac{\left( i - \frac{w-1}{1 + e^{-\omega_u}} \right)^2}{2\sigma^2}} \right]^2 \quad (3.10)$$

### 3.4.2 Estrategia de optimización

La función ECM se puede interpretar como una superficie de error ubicada en un espacio de parámetros  $\mathbb{P}^N$  con  $N=2B_{pv}$ . Esta superficie de error contiene mínimos locales y puntos de silla, por lo cual no es posible asegurar la convergencia de un algoritmo optimizador al mínimo global de la función partiendo de cualquier punto inicial. La presencia de puntos de silla en la función provoca que el uso de algoritmos de optimización basados en la información brindada por la derivada de la función, como el método de gradientes conjugados, no sea una solución al presente problema de optimización. Por este motivo en este trabajo se utiliza el algoritmo de optimización Downhill Simplex, el cual de igual forma no asegura la convergencia al mínimo global de superficies de error con mínimos locales, sin embargo, no es sensible a los puntos de silla ya que la optimización se basa únicamente en la información brindada por la función objetivo.

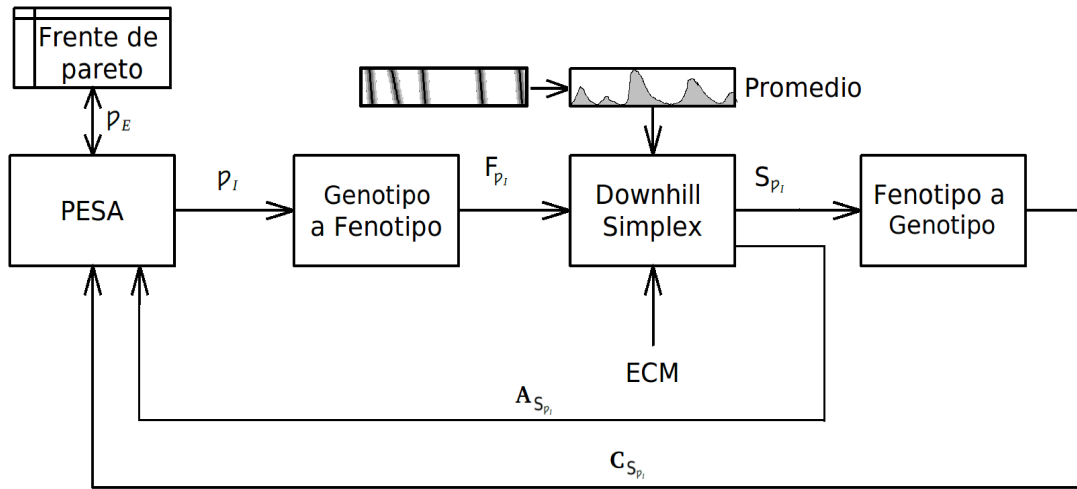
La solución presentada propone optimizar la función objetivo a partir de diferentes puntos ubicados en distintas regiones sobre la superficie de error y así facilitar la búsqueda del mínimo global de la función ECM. Se propone como solución el uso de los algoritmos genéticos como generadores de puntos N-dimensionales a optimizar, en combinación con el optimizador Downhill Simplex para asegurar la ubicación del mínimo de la función ECM o de al menos un punto que aproxime la distribución de intensidad de la ventana evaluada.

Por convención, un punto  $\mathbf{P}$  dentro del espacio  $\mathbb{P}^N$  esta constituido por los parámetros de las bandas de la siguiente manera:

$$\mathbf{P} = [\psi_0 \ \omega_0 \ \psi_1 \ \omega_1 \ \dots \ \psi_{B_{pv}-1} \ \omega_{B_{pv}-1}]^T \quad (3.11)$$

Cada punto  $\mathbf{P}$  es considerado por el algoritmo genético como un *individuo*, donde la representación anterior corresponde al *fenotipo* del mismo y la conversión a una cadena de bits permite obtener su representación como *genotipo*. La cantidad de bits utilizados para representar cada parámetro define la precisión de la conversión de fenotipo a genotipo y viceversa. Como criterio de diseño se utilizaron 24 bits para representar cada parámetro de  $\mathbf{P}$  y se seleccionó un máximo de 8 posibles bandas a ubicar dentro de cada ventana, dando como resultado cromosomas de un largo de 192 bits para representar las 16 dimensiones.

La figura 3.7 ilustra los módulos implementados para la optimización de la función objetivo para el caso de la proyección promedio de las 5 filas centrales de la imagen del carril.



**Figura 3.7:** Diagrama modular de la estrategia de optimización de la función objetivo para el caso de la intensidad promedio

En cada iteración el algoritmo genético PESA genera una población interna  $p_i$  con  $n_i$  nuevos individuos, posteriormente estos son convertidos a su respectivo fenotipo, dando como resultado la población de fenotipos a optimizar  $F_{p_i}$ . Cada fenotipo de esta población es utilizado como punto  $P_0$  o vértice para la creación de un simplex a optimizar, obteniendo los restantes vértices para cada simplex mediante la ecuación:

$$P_i = P_0 + e_{i-1} \quad (3.12)$$

donde  $e$  es de tamaño  $N \times N$  y el subíndice  $i-1$  indica la fila a sumar, la matriz  $e$  utilizada se define como:

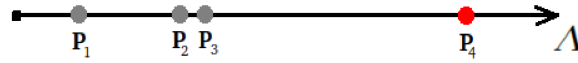
$$e = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \quad (3.13)$$

Los resultantes  $n_i$  simplex son utilizados cada uno como punto de partida por el algoritmo Downhill Simplex para optimizar la función ECM con respecto a la distribución de intensidad de la fila que contiene la proyección promedio de la ventana. Para cada uno de estos simplex el optimizador retorna el mínimo de la superficie de error al cual converge y el error  $\varepsilon$  en ese mínimo, produciendo así  $n_i$  puntos localmente optimizados, que dan origen a la *población de individuos localmente optimizados*  $S_{p_i}$ . Cada individuo de la población  $S_{p_i}$  está relacionado con un factor de *aptitud*, que indica el nivel de ajuste entre la proyección de intensidad promedio de la ventana y la estimada mediante los parámetros que lo conforman, definido matemáticamente como:

$$\Lambda = \frac{1}{1 + \varepsilon} \quad (3.14)$$

Por último la población  $S_{p_i}$  es convertida a su representación en cromosomas y es retornada junto con sus respectivos factores de aptitud al algoritmo PESA, donde el frente de Pareto de la optimización es construido en cada iteración con base en los factores de aptitud de las parametrizaciones de la población  $S_{p_i}$  retornada y la actual  $p_E$ . De esta forma la población interna utilizada en la primera iteración por el algoritmo genético es generada aleatoriamente y las restantes poblaciones internas son generadas a partir de la mutación y el cruce de los individuos localmente optimizados que se encuentran en el actual frente de Pareto.

Para este caso, en que solo se optimiza con referencia a la fila que contiene el promedio de intensidad, el frente de Pareto resultante es de una dimensión ya que solo se tiene una función de aptitud, tal y como es mostrado en la figura 3.8, donde el punto  $P_4$  es el único elemento *no dominado*, es decir  $P_4$  es el frente de Pareto.



**Figura 3.8:** Frente de Pareto para el análisis de una fila promedio.

Terminado el proceso iterativo que da como resultado el frente de Pareto, el elemento que se encuentre en el frente es el punto que optimiza la función ECM y por ende el punto que contiene los parámetros de intensidad y ubicación de las bandas contenidas en la ventana analizada del carril. Finalmente con el objetivo de mejorar la ubicación del mínimo éste es optimizado una última vez con el Downhill Simplex, estableciendo una nueva tolerancia  $\xi_f$  para la condición de parada tal que  $\xi_f < \xi_{FP}$  con  $\xi_{FP}$  la tolerancia utilizada durante la creación del frente de Pareto.

Por otra parte, para el caso en el que se consideran  $p$  filas de la ventana para el análisis, cada individuo de la población  $p_i$  es optimizado por Downhill Simplex  $p$  veces, ya que se optimiza la función ECM con respecto a cada una de las filas en consideración. Esto produce para cada individuo de  $p_i$  un grupo de  $p$  factores de aptitud y  $p$  individuos localmente optimizados que son agrupados en una matriz definida como:

$$Y = \begin{bmatrix} \psi_0 & \omega_0 & \dots & \psi_{B_{pv}-1} & \omega_{B_{pv}-1} \\ \psi_0 & \omega_0 & \dots & \psi_{B_{pv}-1} & \omega_{B_{pv}-1} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \psi_0 & \omega_0 & \dots & \psi_{B_{pv}-1} & \omega_{B_{pv}-1} \end{bmatrix} \quad (3.15)$$

tal que, el punto localmente optimizado que mejor representa al conjunto de las  $p$  filas se obtiene de la mediana  $M_d$  de cada columna de la matriz  $Y$ .

$$P_{sub} = [M_d(Y_{C_0}) \quad M_d(Y_{C_1}) \quad \dots \quad M_d(Y_{C_{N-2}}) \quad M_d(Y_{C_{N-1}})]^T \quad (3.16)$$

Para este caso se obtiene un frente de Pareto de  $p$  dimensiones, ya que cada individuo localmente optimizado tiene un grupo de factores de aptitud  $\Lambda_i \quad i \in \{0, 1, \dots, p-1\}$ , donde el individuo no dominado del frente de Pareto es la parametrización que optimiza la función ECM en las  $p$  filas evaluadas.

Finalmente como criterio de diseño en la solución propuesta se asigna a cualquier fenotipo de la población interna que contenga un parámetro mayor a los rangos posibles para los parámetros de las bandas un factor de aptitud de cero y es retornado como individuo localmente optimizado para evitar su uso como padre de nuevos individuos.

Así el Algoritmo 2 sintetiza la optimización de la función objetivo.

---

#### **Algoritmo 2 para la optimización de la función objetivo**

---

*Entrada:* Ventana del carril  $V_i(x, y)$  de tamaño  $w \times m$

*Salida:* Vector de parámetros de las bandas de la ventana  $PBV$

1. Inicializar la población externa  $p_E$  como un grupo vacío de elementos.
  2. Inicializar la población suboptimizada  $S_{p_i}$  como un grupo vacío de elementos.
  3. Inicializar la población interna  $p_i$  con  $n_i$  individuos aleatorios.
  4. **repetir**
  5.   Encontrar todos los individuos en  $S_{p_i}$  que no son dominados por un elemento de  $S_{p_i} \cup p_E$  e incorporarlos en  $p_E$
  6.   **mientras**  $|p_E| > n_E$  **hacer**
  7.     seleccionar y remover un individuo de  $p_E$
  8.   **fin mientras**
  9.   remover todos los restantes individuos de  $S_{p_i}$  y  $p_i$
  10. **mientras**  $|p_i| < n_i$  **hacer**
  11.   **si** hay probabilidad de cruce  $P_c$  **entonces**
  12.     Seleccionar dos cromosomas padres de  $p_E$  y producir cromosoma hijo mediante cruce y mutación
  13.   **de lo contrario**
  14.     seleccionar un cromosoma padre de  $p_E$  y obtener un cromosoma hijo mediante mutación
  15. **fin si**
-

16. *añadir cromosoma hijo a  $p_i$*
17. **fin mientras**
18. **mientras**  $|S_{p_i}| < n_i$  **hacer**
19. *seleccionar un cromosoma  $C_i$  de  $p_i$  y convertirlo a fenotipo  $F_i$*
20. *si algún parámetro de  $F_i$  está fuera de rango entonces*
21. *añadir  $C_i$  con  $\Lambda=0$  a  $S_{p_i}$*
22. **de lo contrario**
23. *crear simplex con el fenotipo*
24. **si** Analizar fila promedio **entonces**
25.  $P_{sub} = \text{Optimizar Simplex con respecto a la fila promedio con tolerancia } \xi_{FP}$   

$$\Lambda_0 = \frac{1}{1 + \varepsilon}$$
26. **de lo contrario**
27.  $k = \text{fila inicial}$
28. **mientras**  $k \neq \text{fila final}$
29.  $P_{sub_k} = \text{Optimizar Simplex con respecto a la fila } k \text{ con tolerancia } \xi_{FP}$   

$$\Lambda_k = \frac{1}{1 + \varepsilon_k}$$
30. *añadir  $P_{sub_k}$  a  $Y$*
31.  $k = \text{nueva fila}$
32. **fin mientras**
33.  $P_{sub} = [M_d(Y_{C_0}) \ M_d(Y_{C_1}) \ \dots \ M_d(Y_{C_{N-2}}) \ M_d(Y_{C_{N-1}})]$
34. **fin**
35. *añadir  $P_{sub}$  con su aptitud(es)  $\Lambda$  a  $S_{p_i}$*
36. **fin**
37. **hasta** alcanzar máximo número de iteraciones
38.  $PBV = \text{individuo no dominado del frente de Pareto}$
39. **si** Analizar fila promedio **entonces**
40.  $PBV = \text{Optimizar Simplex con respecto a la fila promedio con tolerancia } \xi_f$
41. **retornar**  $PBV$

## 3.5 Procesamiento de las bandas encontradas

En esta sección se detallan las estrategias utilizadas para remover las bandas falsas o duplicadas debido a ruido en la imagen de carril y al traslape de ventanas analizadas.

### 3.5.1 Umbralización y organización de bandas

El vector de parámetros de bandas encontradas en la ventana  $PBV$  es sometido a segmentación por umbralización de acuerdo al parámetro  $A$  estimado para cada banda encontrada y a un valor de intensidad umbral  $T$  establecido como parámetro del algoritmo, y simultáneamente

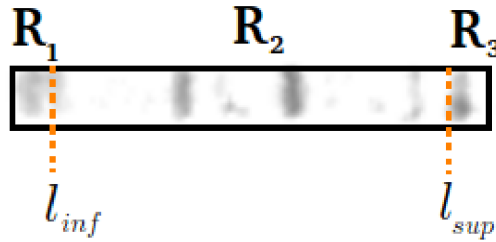
ordenado ascendentemente de acuerdo al parámetro  $\mu$  de cada banda. Las bandas con una intensidad máxima menor al umbral, 20 unidades de intensidad como criterio de diseño para la solución propuesta en el presente trabajo, son eliminadas de  $PBV$  y en caso contrario son ordenadas, esto es:

$$PBV = \begin{cases} \text{ordenar según } \mu_u & \text{si } A_u \geq T \\ \text{eliminar banda} & \text{en caso contrario} \end{cases}$$

### 3.5.2 Corrección del duplicado de bandas entre ventanas

Debido al traslape de  $8\sigma$  existente entre cada par de ventanas adyacentes, las bandas presentes en la región de traslape pueden ser duplicadas por el análisis, con valores de  $A$  y  $\mu$  diferentes en algunos decimales, dependiendo de la precisión de las variables utilizadas para representarlas y de si el punto mínimo de ECM encontrado por el algoritmo para cada ventana las incluye o no. Para evitar el duplicado de bandas en el vector final se propone el algoritmo explicado en esta sección.

La ventana analizada  $V_i$  se divide en tres regiones como se ilustra en la figura 3.9, donde las regiones  $R_1$  y  $R_3$  corresponden al área de traslape de la actual ventana con las ventanas adyacentes  $V_{i-1}$  y  $V_{i+1}$ .



**Figura 3.9:** División de la ventana para corregir duplicado de bandas

Los límites de las regiones son calculados como:

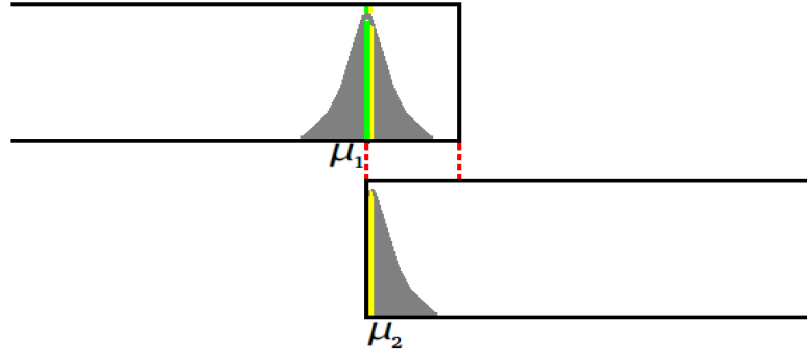
$$\begin{aligned} l_{inf} &= 8\sigma \\ l_{sup} &= (w-1) - 8\sigma \end{aligned} \quad (3.17)$$

En el vector de parámetros de bandas de la ventana  $PBV$  se buscan las bandas que se encuentran en las regiones  $R_1$  y  $R_3$  esto según el parámetro  $\mu$  de la banda y son almacenadas en un vector de acuerdo a la región en la que se encuentran:

$$PBV = \begin{cases} \text{Almacenar en } V_{R_1} & \text{si } \mu_u \leq l_{inf} \\ \text{Almacenar en } V_{R_3} & \text{si } \mu_u \geq l_{sup} \end{cases}$$

Se calcula el valor de error ECM que tiene el vector  $V_{R_1}$  con respecto a la región del carril

$R_1$  y se compara con el error que presentan las bandas ubicadas durante la iteración anterior en la región  $R_3$  donde  $R_{1_{V_i}} = R_{3_{V_{i-1}}}$ , tal que, si el error de las actuales bandas de  $V_{R_1}$  es menor, las bandas ubicadas en la región  $R_3$  de la ventana anterior son sustituidas copiando completamente el vector  $PBV$  en el vector  $\Theta$  que contiene los parámetros de todas las bandas encontradas en el carril hasta el momento. Cuando esta condición se cumple por que en la ventana anterior  $V_{i-1}$  en su región  $R_3$  no se encontraron bandas, se evalúa el caso ilustrado en la figura 3.10, en el cual el vector  $PBV$  contiene una banda en el primer píxel de la ventana actual y en la ventana anterior se encontró una banda un píxel antes, en este caso la primera banda de  $PBV$  se elimina ya que fue encontrada considerando menos información que la de la ventana anterior



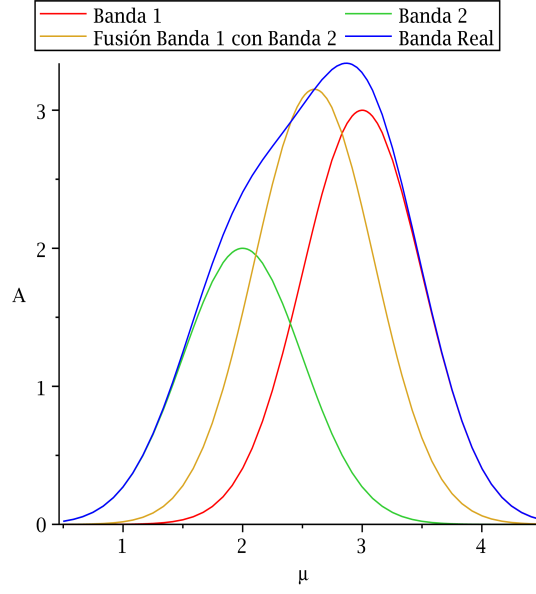
**Figura 3.10:** Duplicado de bandas en los límites de la ventana

Por otra parte si la actual estimación de la distribución de intensidad de  $R_1$  presenta mayor error se almacenan en  $\Theta$  solo las bandas de  $PBV$  con un  $\mu > I_{inf}$ . Por último se calcula el error ECM del vector  $V_{R_3}$  con la región  $R_3$  para ser comparado en la siguiente iteración.

### 3.5.3 Corrección del duplicado de bandas en el vector de parámetros final

Debido al ruido presente en la imagen del gel y a que el perfil de intensidad real de las bandas no es perfectamente gaussiano, es posible que el método establezca varias bandas de diferentes amplitudes con valores medios diferenciados por uno o dos píxeles para estimar con menor error el perfil de intensidad real de la banda, esta situación es ilustrada en la figura 3.11, en la cual la *Banda 1* y la *Banda 2* han sido seleccionadas por el algoritmo debido a que su combinación modela la distribución de intensidad de la *banda real*.





**Figura 3.11:** Aproximaci3n de una banda real mediante varias bandas estimadas

Para corregir este efecto se desarrolla un algoritmo que recorre el vector de parámetros final  $\Theta$  en b3squeda de bandas que pueda fusionar, utilizando como criterio que la distancia existente entre dos bandas debe ser mayor a un parámetro  $\Delta \mu$  establecido manualmente. De lo contrario las bandas analizadas son fusionadas y la nueva banda es comparada con la siguiente dentro del vector  $\Theta$ , recordando que el vector de parámetros finales ha sido previamente ordenado ascendentemente esto es:

$$\Theta = \begin{cases} \text{Fusionar bandas} & \text{si } \mu_{u+1} - \mu_u \leq \Delta \mu \\ \text{Almacenar } \mu_u & \text{si } \mu_{u+1} - \mu_u > \Delta \mu \end{cases}$$

la posici3n central de la nueva banda es determinada mediante [28]:

$$\mu_t = \frac{\mu_u A_u + \mu_{u+1} A_{u+1}}{A_u + A_{u+1}} \quad (3.18)$$

y la intensidad m3xima de la banda es leída directamente de la imagen del carril en el píxel  $\mu_t$ .

# Capítulo 4

## 4 Análisis de Resultados

En este capítulo se presentan los resultados obtenidos para cada uno de los módulos desarrollados e implementados en el presente trabajo para la detección automática de bandas en las imágenes de geles de electroforesis. Inicialmente se realiza un análisis en etapas de los resultados obtenidos, divididas según el módulo bajo evaluación y finalmente se exponen los resultados producto del funcionamiento del sistema completo. Para realizar las pruebas se utilizó una computadora con dieciséis procesadores Intel Xeon X5570 con una frecuencia de 2.93GHz y 8GB de RAM.

Lo restante del presente capítulo se divide en las siguientes secciones:

- Estimación de  $\sigma$
- Extracción del fondo según  $\sigma$
- Optimización de la función objetivo
- Umbralización
- Corrección del duplicado de bandas en las regiones de traslape
- Fusión de bandas.
- Medición de la desviación promedio de las bandas estimadas.
- Análisis de carriles de geles de electroforesis
- Rendimiento

El análisis expuesto involucra pruebas realizadas al algoritmo tanto con carriles *reales* extraídos de geles, como con carriles *sintéticos*, es decir, carriles creados por medio de algoritmos, con una cantidad de bandas establecida manualmente y distribuidas aleatoriamente a lo largo del carril. De igual forma sucede con la intensidad de las bandas, las cuales son generadas aleatoriamente dentro de un

rango establecido. Para estos carriles sintéticos se conoce el vector de parámetros que dio origen al carril, permitiendo así realizar evaluaciones del rendimiento del método propuesto, considerando la diferencia entre la cantidad de bandas en el carril y la cantidad de bandas estimadas, así como la desviación promedio entre la ubicación de las bandas originales y la ubicación estimada y el error total existente entre la distribución de intensidad estimada y la real.

## 4.1 Estimación de $\sigma$

Para las pruebas del módulo de estimación de  $\sigma$  se crean carriles sintéticos con 10 valores diferentes de desviación estándar para las bandas presentes en los carriles. Los valores son seleccionados de tal forma que el ancho resultante de la banda, aproximadamente  $(6\sigma)$ , presente en el carril sintético sea un valor entero en píxeles. Para cada valor de desviación estándar se generan 10 carriles con 45 bandas distribuidas aleatoriamente. Estos son utilizados como entrada al módulo de estimación y se obtienen las estimaciones de  $\sigma$  para las bandas presentes en cada carril, los resultados son mostrados en la tabla 4.1

**Tabla 4.1:** Valores reales y estimados de  $\sigma$  para carriles sintéticos

Número de Medición	Valores de $\sigma$ (píxeles)									
	0,5	0,67	0,83	1	1,17	1,33	1,5	1,67	1,83	2
1	0,59	0,79	0,57	0,94	1,15	1,78	1,15	0,71	0,33	0,89
2	0,48	0,59	0,71	1,01	0,71	0,71	0,94	0,71	0,89	0,43
3	0,43	0,79	0,84	0,53	1,42	1,78	1,64	1,32	0,59	0,68
4	0,49	0,71	0,65	0,94	0,75	1,32	1,42	1,64	0,59	0,43
5	0,45	0,49	0,89	0,84	1,32	1,64	1,64	1,23	0,33	0,65
6	0,45	0,57	0,59	1,23	1,23	0,59	1,64	0,59	1,92	0,55
7	0,68	0,71	0,68	0,75	1,15	1,23	1,42	0,59	1,32	1,92
8	0,43	0,71	0,71	1,01	1,42	0,71	0,65	0,75	0,71	0,94
9	0,59	0,59	0,59	0,75	0,94	0,59	1,15	1,78	1,01	0,33
10	0,57	0,71	0,84	0,75	1,07	1,78	0,33	1,78	0,33	0,59

Para realizar una interpretación de los datos obtenidos se calcula la desviación estándar (SD) y el coeficiente de dispersión (CV) del total de valores estimados para cada valor de  $\sigma$  utilizado en la creación de carriles sintéticos. Para un coeficiente de dispersión definido como:

$$CV = \frac{SD * 100}{\bar{\sigma}} \quad (4.1)$$

se obtienen los valores mostrados en la tabla 4.2

**Tabla 4.2:** Medidas estadísticas para las estimaciones de  $\sigma$  de los carriles sintéticos

$\sigma$ original (píxeles)	0,5	0,67	0,83	1	1,17	1,33	1,5	1,67	1,83	2
$\sigma$ medio estimado (píxeles)	0,51	0,67	0,71	0,88	1,12	1,21	1,2	1,11	0,8	0,74
SD (píxeles)	0,08	0,1	0,11	0,19	0,25	0,52	0,45	0,5	0,51	0,46
CV(%)	16,18	15	16,16	22,11	22,42	42,71	37,45	44,6	63,21	61,16

Se observa que el algoritmo realiza estimaciones con mayor precisión para los valores de  $\sigma$  menores a 1,17 píxeles ya que para estos casos se obtiene una desviación estándar máxima de 0,25 píxeles y un coeficiente de variación de 22,42% máximo; sin embargo, para valores de  $\sigma$  mayores a 1,17 píxeles se obtiene un coeficiente de variación de hasta un 63%. Esto se debe a que el método de estimación es un proceso iterativo que utiliza en cada iteración una distribución de intensidad diferente como referencia, ya que esta varía según el tamaño del kernel utilizado para extraer el fondo de la ventana con mayor cantidad de bandas, que a su vez depende del  $\sigma$  evaluado en cada iteración. Debido a que la extracción del fondo se basa en la apertura morfológica, para valores pequeños del kernel si las bandas contenidas en el carril son de tamaño mayor al kernel éste las *erosiona* dando como resultado bandas más delgadas y aunque el error obtenido de la estimación por píxel sea mayor con el actual  $\sigma$  que cuando se está encontrando la aptitud con el valor real, el error del conjunto o de la banda completa es menor para la banda más delgada que para la banda real, que contiene más píxeles y por ende mayor error acumulativo y menor aptitud. La tabla 4.3 muestra como para las bandas más anchas el método pierde exactitud en la estimación de  $\sigma$ .

**Tabla 4.3:** Anchos originales y estimados de las bandas de los carriles sintéticos

Ancho de la banda original (píxeles)	3	4	5	6	7	8	9	10	11	12
Ancho de la banda estimada (píxeles)	3,06	4,01	4,24	5,25	6,69	7,28	7,19	6,66	4,82	4,45

Debido a esta limitante las restantes pruebas en el presente capítulo se realizan estableciendo manualmente el  $\sigma$  de las bandas presentes en el carril.

## 4.2 Extracción del fondo según $\sigma$

Para esta evaluación se realiza la extracción del fondo de una imagen de un carril sintético y de un carril extraído de la imagen de un gel, asumiendo que el valor de la desviación estándar que define el perfil de las bandas contenidas en cada uno es conocido. Para el caso del carril real este se determina directamente de la imagen, conociendo que para una banda  $6\sigma=99,74\%$  de su ancho. Para ambas pruebas se asigna un valor de 12 al factor de proporción para el tamaño del kernel a utilizar para remover el fondo ( $\varphi=12$ ).

Inicialmente se elimina el fondo de un carril sintético de tamaño  $400 \times 20$  con 10 bandas caracterizadas con  $\sigma=1,33$  píxeles, el cual se ilustra en la figura 4.1. Los parámetros que describen a las bandas son conocidos, así como el valor del fondo del carril el cual se establece en 70 unidades de intensidad  $ui$ .



**Figura 4.1:** Carril sintético utilizado para evaluar la extracción del fondo

En el carril resultante sin fondo mostrado en la figura 4.2 se mide para la fila central el valor de intensidad  $A_u$  en las posiciones  $\mu_u$  correspondientes a las ubicaciones de las bandas que dieron origen al carril con fondo, dando como resultado los valores mostrados en la tabla 4.4.



**Figura 4.2:** Carril sintético sin fondo resultante

En adelante se utiliza como convención el nivel 0 para la máxima intensidad y 255 para la mínima.

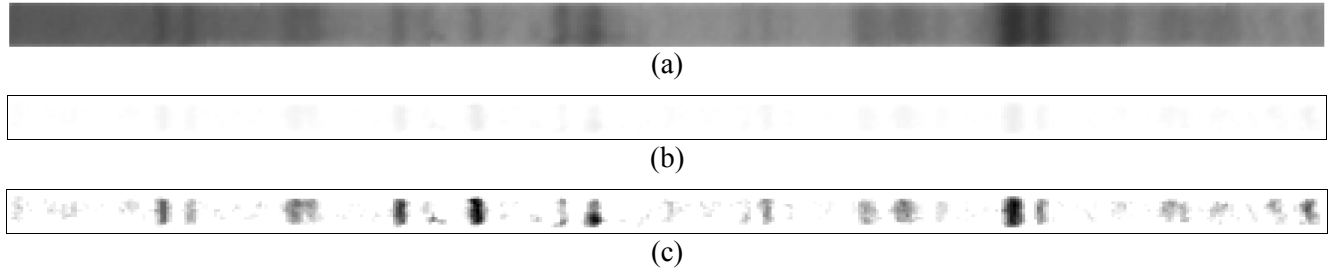
**Tabla 4.4:** Valores de intensidad para las bandas en un carril sintético con fondo y sin fondo

Intensidad ( $ui$ )	Ubicación central $\mu$ (píxeles)									
	18	39	50	144	162	215	240	268	351	366
$A_{fondo}$	255	163	255	255	253	255	130	216	182	190
$A_{sinfondo}$	185	92	184	185	183	185	60	146	112	120
$A_{fondo} - A_{sinfondo}$	70	71	71	70	70	70	70	70	70	70

Se observa que para el caso del carril sintético solo es eliminado del carril lo considerado como fondo, el cual también es eliminado de la amplitud máxima de las bandas.

Para la extracción del fondo del carril real se utiliza el carril de la figura 4.3a, el cual contiene bandas con  $\sigma=1$  píxel, dando como resultado el carril mostrado en la figura 4.3b. En este caso el carril

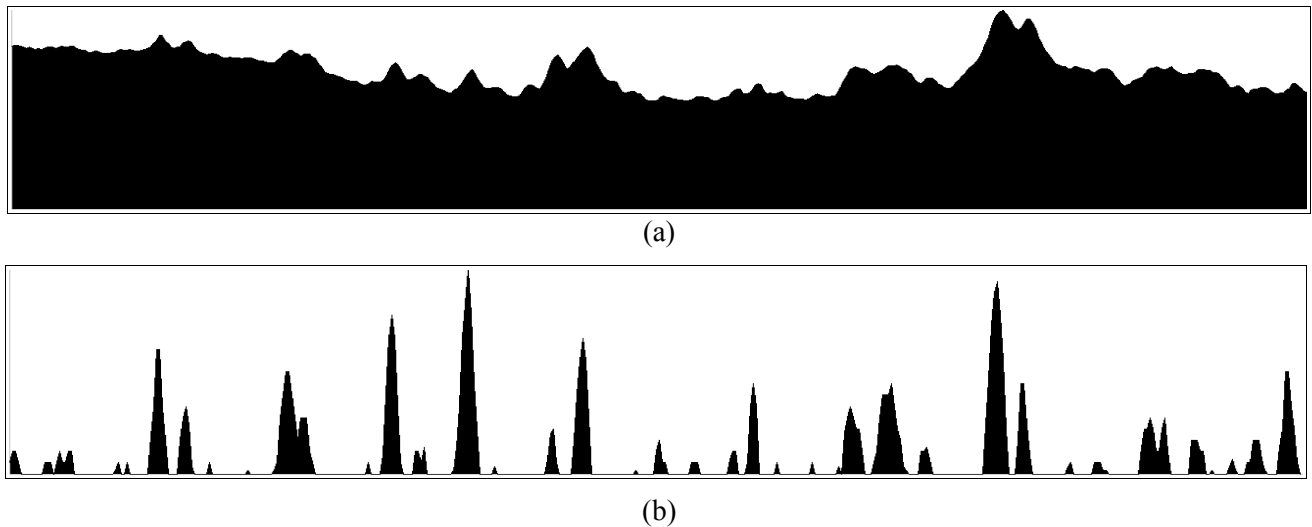
resultante contiene las bandas presentes más un ruido atenuado.



**Figura 4.3:** Extracción del Fondo. (a)Carril real, (b)Carril real sin fondo y (c) Carril sin fondo con mejora de contraste.

Como se observa en la figura 4.3b, el carril resultante presenta niveles de intensidad altos en comparación con los niveles de intensidad presentes en el carril original, esto se debe al bajo contraste de las imágenes de geles de electroforesis. La figura 4.3c muestra el carril resultante al aplicar la mejora de contraste propuesta en [4].

Por otra parte en la figura 4.4 se muestra la distribución de intensidad de la fila central del carril original y la distribución de intensidad del carril sin fondo. Es posible observar como el método elimina el fondo de la imagen y reduce el ruido, permitiendo obtener un carril con una distribución de intensidad más descriptiva del perfil de las bandas, evitando así que alguna de las funciones gaussianas de la función objetivo sea ajustada al fondo de la imagen.



**Figura 4.4:** Distribución de intensidad de la fila central del carril. (a) Carril con fondo y (b) Carril sin fondo

### 4.3 Optimización de la función objetivo

Para evaluar el funcionamiento del optimizador de la función objetivo se utiliza la ventana de un carril sintético, mostrada en la figura 4.5, de tamaño  $100 \times 20$ , con 6 bandas caracterizadas por un perfil de intensidad con  $\sigma=1$  píxel.



**Figura 4.5:** Carril sintético para la evaluación de la optimización de la función objetivo.

Se evalúa el comportamiento del optimizador para los siguientes casos:

- Cantidad de bandas en la imagen del carril igual a la cantidad de bandas utilizadas por la función objetivo.
- Cantidad de bandas en la imagen del carril mayor a la cantidad de bandas utilizadas por la función objetivo.
- Cantidad de bandas en la imagen del carril menor a la cantidad de bandas utilizadas por la función objetivo.

Por último, se realiza la optimización de la función objetivo utilizando como referencia una ventana de tamaño  $61 \times 13$  de un carril extraído de los geles. En la tabla 4.5 se muestran los parámetros utilizados por el Downhill Simplex y el algoritmo genético para la optimización.

**Tabla 4.5:** Parámetros de PESA y Downhill simplex para la optimización

Parámetro	Valor
PESA	
• Tasa de Mutación inicial	0.5
• Tasa de Mutación final	$1/L^*$
• Poblacion externa	100
• Población interna	6
• Iteraciones	100
Downhill Simplex	
• Tolerancia	0.001
• Cantidad máxima de iteraciones	10000

\* $L$ =cantidad de bits utilizados para representar el cromosoma.

#### 4.3.1 Cantidad de bandas en la imagen del carril igual a la cantidad de bandas utilizadas por la función objetivo

Para este caso se optimiza la función objetivo cinco veces con respecto al carril de la figura 4.5, y se mide la variación existente entre los parámetros estimados  $A$  y  $\mu$  de cada banda con respecto a los valores reales, dando como resultado los valores mostrados en la tabla 4.6

**Tabla 4.6:** Valores reales y estimados de los parámetros de las bandas, considerando todas las bandas presentes en la ventana.

Parámetro	Valor Real	Valores Estimados				
		1	2	3	4	5
$A_0$	234	253,818	255	255	218,965	255
$A_1$	204	208,566	208,587	0	211,63	208,587
$A_2$	185	0,00346	0	201,623	174,38	201,623
$A_3$	241	69,616	255	255	217,442	255
$A_4$	255	79,4338	255	255	224,859	255
$A_5$	255	254,999	255	255	223,797	255
$\mu_0$	16	16,4345	16,4345	16,4345	16,4292	16,4345
$\mu_1$	43	42,7881	42,7881	3,26122	42,7834	42,7881
$\mu_2$	62	62,2	0	62,4473	62,4473	62,4473
$\mu_3$	70	69,616	69,616	69,16	69,6129	69,616
$\mu_4$	79	79,4338	79,4339	79,4339	79,431	79,4339
$\mu_5$	86	86,2464	86,2465	86,2465	86,2425	86,2425
Bandas acertadas (%)	NA*	100%	83,33%	83,33%	100%	100%

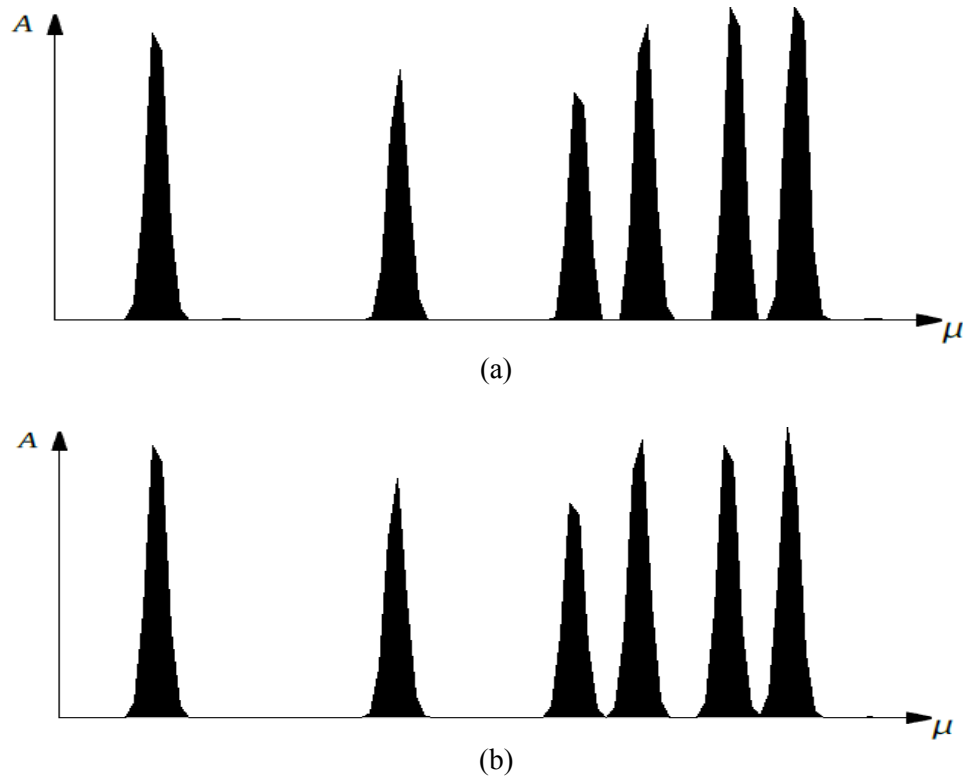
\*No aplica

En los resultados anteriores se considera que se ha acertado con la ubicación de la banda si el valor estimado para el parámetro  $\mu$  de la banda se encuentra a una distancia de 2 píxeles del valor real. Se obtiene como porcentaje mínimo de bandas acertadas un 83,33% lo cual indica que el método de estimación presenta resultados satisfactorios, sin importar el punto del cual se da inició la optimización ya que para cada medición de la tabla 4.6 se utilizan simplex iniciales diferentes. Es posible aumentar el porcentaje de bandas acertadas incrementando la cantidad de iteraciones realizadas por el algoritmo genético, para este caso se utilizan 100 iteraciones.

La figura 4.6 ilustra la distribución de intensidad real y estimada para la medición número 5



presente en la tabla 4.6, se obtiene que el método logra ajustar la función objetivo a la distribución de intensidad utilizada como referencia.



**Figura 4.6:** Distribuciones de intensidad considerando una cantidad de bandas igual a las de la ventana. (a) Distribución del carril, (b) Distribución estimada

#### 4.3.2 Cantidad de bandas en la imagen del carril mayor a la cantidad de bandas utilizadas por la función objetivo

Para evaluar el comportamiento del optimizador en el caso en el cual en la ventana existen mayor cantidad de bandas que las consideradas por la función objetivo, se realiza la optimización con respecto al carril de la figura 4.5 y se establece en la función objetivo una cantidad máxima de bandas a encontrar igual a 5. Los resultados obtenidos son mostrados en la tabla 4.7.

**Tabla 4.7:** Valores reales y estimados de intensidad y ubicación para la optimización que considera menor cantidad de bandas

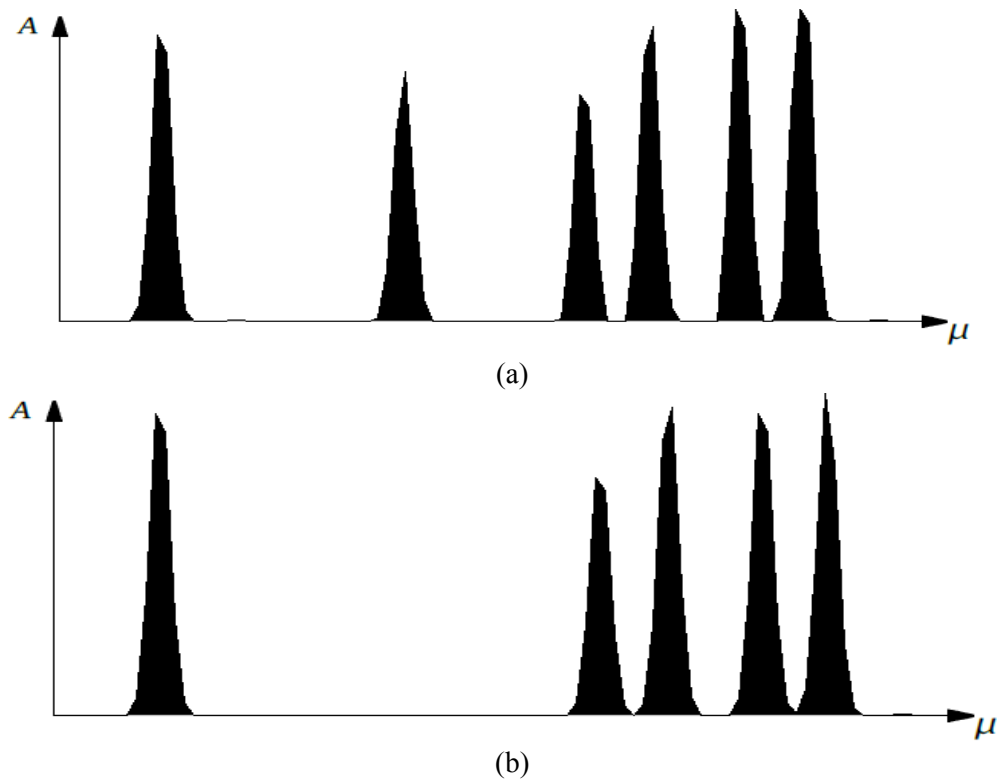
Valor	Intensidad (ui)						Ubicación central (píxeles)					
	$A_0$	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	$\mu_0$	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$	$\mu_5$
Real	234	204	185	241	255	255	16	43	62	70	79	86
Estimado	255	NC*	201,623	255	255	255	16,4345	NC*	62,4473	69,616	79,4339	86,2465

\*No considerada por el optimizador

Se obtiene que el optimizador ubica con exactitud las bandas consideradas por la función objetivo, sin importar la presencia de mayor cantidad de bandas en la ventana. De igual manera se estiman valores de intensidad mayores a los reales, esto es una consecuencia de la mejora de contraste realizada a la ventana antes de realizar la optimización; sin embargo debido a la exactitud con la que se realiza la estimación de  $\mu$  es posible leer el valor máximo real directamente desde la imagen del carril.

Para este caso se obtiene un ECM de  $842\mu\text{i}^2$  con respecto a la proyección promedio de las cinco filas centrales del carril de la figura 4.5, esto a pesar de que la distribución de intensidad es aproximada satisfactoriamente para las restantes bandas.

La figura 4.7 ilustra la distribución real y la estimada que describen los valores de la tabla 4.7.



**Figura 4.7:** Distribuciones de intensidad, mayor cantidad de bandas en la ventana que las consideradas. (a) real, (b) Estimada

#### 4.3.3 Cantidad de bandas en la imagen del carril menor a la cantidad de bandas utilizadas por la función objetivo

Para las ventanas en las cuales se presenta esta situación se observaron dos fenómenos distintos durante la optimización. Para el primero de ellos el optimizador reduce la intensidad  $A$  de una de las bandas estimadas a un valor cercano a cero, de tal forma que la etapa posterior en la cual se realiza la

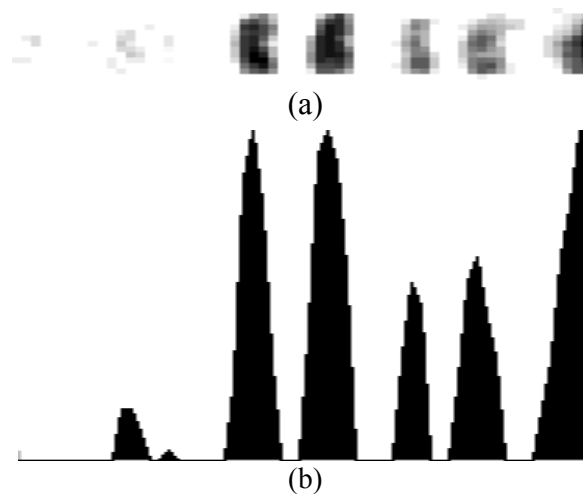
umbralización de bandas encontradas, permite eliminar estas falsas bandas. Para las pruebas se utiliza nuevamente el carril sintético de la figura 4.5 y se optimiza la función de error considerando 7 bandas, dando como resultado los parámetros de bandas contenidos en la tabla 4.8

**Tabla 4.8:** Valores reales y estimados de intensidad y ubicación para la optimización de un carril sintético que considera mayor cantidad de bandas

Valor	intensidad(ui)							Ubicación central(píxeles)						
	$A_0$	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	$A_6$	$\mu_0$	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$	$\mu_5$	$\mu_6$
Real	234	204	185	241	255	255	NE*	16	43	62	70	79	86	NE*
Estimado	254,7	208,7	201,6	254,8	255	255	1.5e-99	16,4	42,8	62,4	69,6	79,4	86,3	11,4

\*No existe en la distribución de intensidad real.

Por otra parte para el caso de los carriles extraídos de los geles de electroforesis, debido a que el perfil de una banda no es completamente gaussiano y al ruido presente en la imagen, el optimizador es capaz de ubicar varias bandas de tal forma que su combinación aproxime con menor error ECM la distribución real de intensidad, sin embargo en la distribución de intensidad solo se observa una banda. Un ejemplo de esto se obtiene al optimizar la función objetivo con respecto a la distribución de intensidad promedio de las cinco filas centrales del segmento de carril de tamaño  $61 \times 13$  mostrado en la figura 4.8a.



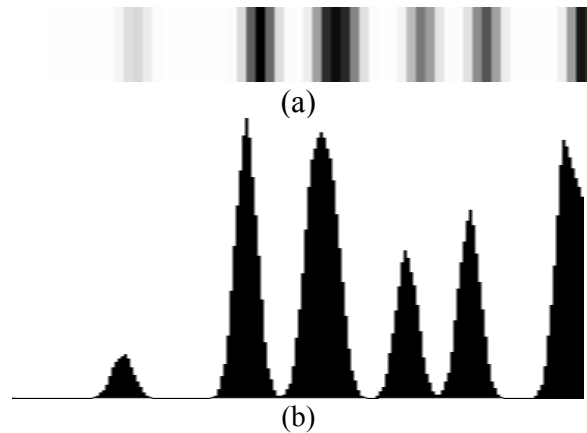
**Figura 4.8:** Carril de referencia. (a) Imagen del carril (b) Distribución de intensidad promedio

Para la optimización se consideran 7 bandas en la función objetivo, y un  $\sigma$  igual a 1, dando como resultado los valores mostrados en la tabla 4.9 para los parámetros de las bandas estimadas.

**Tabla 4.9:** Parámetros de bandas estimados para un carril real con menos bandas que las consideradas por la función objetivo

Parámetro	Banda						
	1	2	3	4	5	6	7
$A(ui)$	40,91	255	199,26	173,50	135,92	171,8	250,87
$\mu(\text{píxeles})$	11,72	24,97	32,23	34,11	42,17	48,89	59,37

Con los valores estimados se crea un carril sintético con el objetivo de observar el grado de similitud entre la distribución de intensidad del carril original y el carril estimado. El carril obtenido es ilustrado en la figura 4.9, en la cual se observa que a pesar de que el método duplica bandas, en el caso de la banda 3 y 4 de la tabla 4.9, la distribución de intensidad total descrita por el conjunto de parámetros se asemeja a la original, confirmando el valor de ECM obtenido el cual fue de  $374,93 ui^2$ .

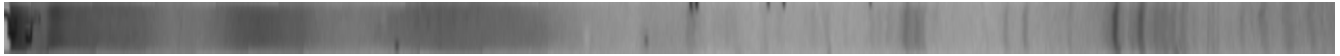


**Figura 4.9:** Carril estimado. (a) Imagen del carril  
(b) Distribución de intensidad estimada

El problema de duplicación de bandas de los resultados obtenidos en esta sección es resuelto por el módulo de fusión de bandas evaluado en las secciones posteriores.

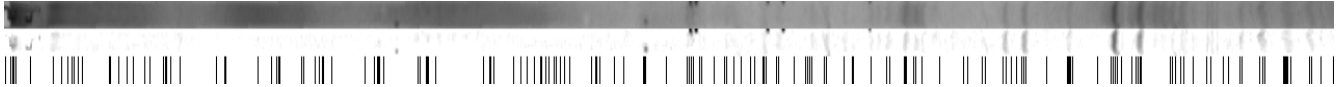
## 4.4 Umbralización

La umbralización del vector de parámetros resultante de la optimización de cada ventana analizada, permite eliminar las funciones gaussianas que fueron ajustas por el optimizador al ruido atenuado que permanece en la distribución de intensidad del carril después de eliminar lo considerado como fondo, esto generalmente para el caso en el cual la ventana analizada contiene una menor cantidad de bandas que las consideradas por la función objetivo. Para evaluar el desempeño del algoritmo de optimización con umbralización se utiliza el carril de la figura 4.10, ya que presenta grandes regiones sin bandas permitiendo observar con mayor facilidad la función de la umbralización.



**Figura 4.10:** Carril utilizado para pruebas de umbralización

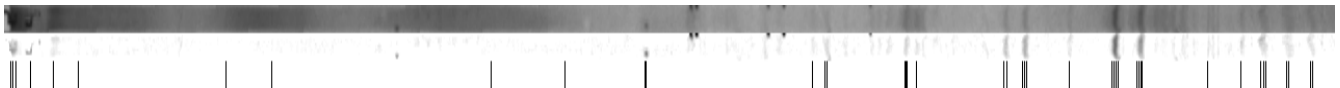
Primeramente se realiza la optimización del carril sin considerar el módulo de umbralización y se genera un carril sintético con los parámetros de bandas estimados resultantes. Para cada banda únicamente se señala su posición central en el carril, permitiendo observar con mayor facilidad la cantidad de bandas individuales presentes, dando como resultado el carril ilustrado en la figura 4.11, en la cual además se muestra el carril original y el carril original sin fondo considerado para la estimación.



**Figura 4.11:** Estimación de la ubicación central de bandas sin utilizar umbralización

Se puede observar que en la primera mitad del carril no deberían existir bandas, sin embargo el método las ubica durante el análisis de la respectiva ventana con una intensidad similar a la del fondo que aún quedó en la imagen. Para este caso el vector de parámetros final contiene 200 bandas.

Se realiza de nuevo la optimización del carril pero considerando un valor de umbral para las bandas de 20ui como criterio de diseño, dando como resultado el carril con ubicaciones centrales de bandas ilustrado en la figura 4.12.



**Figura 4.12:** Estimación de la ubicación central de bandas con umbralización

De esta forma se obtienen un total de 42 posibles bandas utilizando la umbralización, permitiendo eliminar el ruido del vector de parámetros de bandas final.

## 4.5 Corrección del duplicado de bandas en las regiones de traslape

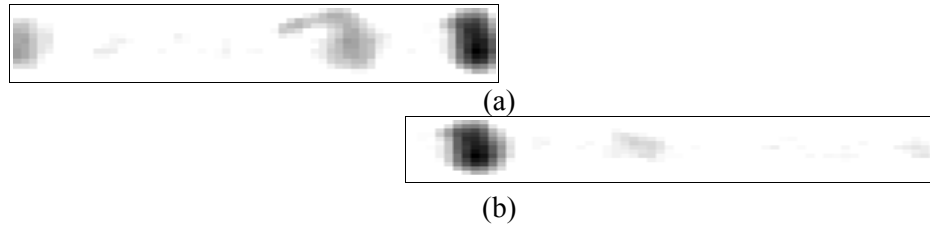
Para evaluar la corrección del duplicado de bandas debido al traslape de dos ventanas se selecciona como imagen de referencia el segmento de carril mostrado en la figura 4.13, de tamaño

$110 \times 13$  y se optimiza utilizando la proyección promedio de las 5 filas centrales del carril, en un análisis en ventanas de tamaño  $w=60$  píxeles y con 8 bandas consideradas por ventana, para un valor de  $\sigma=1,13$  píxeles.



**Figura 4.13:** Segmento de carril utilizado para la evaluación del duplicado de bandas debido al traslape

De tal forma el análisis se divide en dos ventanas con el fondo previamente extraído, ilustradas en la figura 4.14



**Figura 4.14:** Segmentación en ventanas con duplicado de bandas en la region de traslape (a) Primera ventana (b) Segunda ventana

Se realiza inicialmente la optimización de la función objetivo sin utilizar la corrección de traslape propuesta en el presente trabajo dando como resultado las ubicaciones centrales para cada banda mostradas en la tabla 4.10

**Tabla 4.10:** Ubicación central de las bandas estimadas sin el algoritmo de corrección del traslape

Ubicación central	Bandas										
	1	2	3	4	5	6	7	8	9	10	11
$\mu$	1.84e-101	39,30	41,4	43,59	55,46	57,57	60	55,39	57,34	59,22	76,5

Posteriormente se realiza de nuevo la optimización considerando el método propuesto para la corrección del duplicado de bandas, obteniendo los parámetros mostrados en la tabla 4.11

**Tabla 4.11:** Ubicación central de las bandas estimadas con el algoritmo de corrección del traslape

Ubicación central	Bandas								
	1	2	3	4	5	6	7	8	9
$\mu$	1.856e-101	39,30	41,4	43,59	55,46	57,57	59,22	60	76,5

Debido a que el  $\sigma$  utilizado es de 1,13 píxeles la región de traslape de ventanas va del píxel 51 al 60, región en la cual para el vector de parámetros obtenido con el algoritmo de corrección las bandas 5 y 6 de la tabla 4.10 se prefirieron ante las bandas 8 y 9 de la misma tabla. Esto debido a que las bandas 5 y 6 de la tabla 4.10 presentaron un ECM de  $5,63ui^2$ , mientras que las bandas 8 y 9 un ECM de  $1379,48ui^2$ .

## 4.6 Fusión de bandas

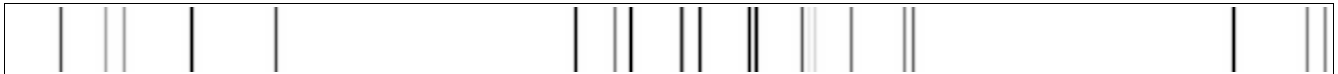
Para la evaluación del algoritmo de fusión de bandas se utiliza el carril sintético presente en la figura 4.15.



**Figura 4.15:** Carril sintético para la evaluación de la fusión de bandas

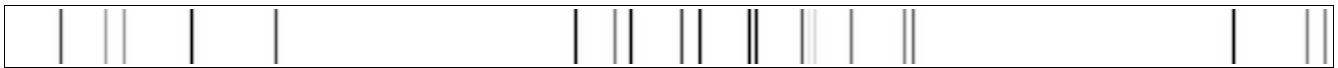
Este carril contiene 20 bandas con una desviación estándar de 1,1 píxeles para cada banda. Se realiza la optimización de la función objetivo con respecto al carril y se mide la variación del error entre la cantidad de bandas estimadas y las reales, esto al considerar la fusión. A su vez se observa el efecto de la fusión de bandas sobre la distribución de intensidad del carril resultante.

Inicialmente se realiza la optimización sin considerar la fusión de bandas, dando como resultado la ubicación de 26 bandas, las cuales son utilizadas para crear el carril ilustrado en la figura 4.16. Se observa que a pesar de existir un error de un 30% en la cantidad de bandas estimadas la distribución de intensidad total del conjunto se ajusta a la distribución del carril original, esto se debe a que varias bandas son combinadas para formar una sola que se ajuste a la de referencia, causando la duplicación de bandas con diferencias típicas de 1 ó 2 píxeles para su parámetro  $\mu$ .



**Figura 4.16:** Carril estimado sin considerar la fusión de bandas

Por otra parte se realiza la optimización considerando el algoritmo de fusión de bandas y utilizando un criterio de fusión  $\Delta\mu=2$  píxeles. Se obtienen como resultado los parámetros de 21 bandas para describir la distribución de intensidad de la figura 4.15. Estos fueron utilizados para crear el carril sintético ilustrado en la figura 4.17



**Figura 4.17:** Carril estimado considerando la fusión de bandas

De esta forma se reduce el error entre la cantidad real y estimada a un 5%, manteniendo el ajuste realizado por el optimizador con el carril real.

## 4.7 Medición de la desviación promedio de las bandas estimadas.

Para realizar la medición de la desviación promedio de las bandas estimadas se crea un carril  $S_c$  de tamaño  $700 \times 20$  con 25 bandas con desviación estándar de 1,13 píxeles, y se utiliza el algoritmo propuesto para optimizar la función objetivo ECM, considerando ventanas de 60 píxeles y un

máximo de 8 bandas por ventana, con un valor de umbral establecido en  $20\mu_i$  y un criterio de fusión  $\Delta\mu=2$  píxeles.

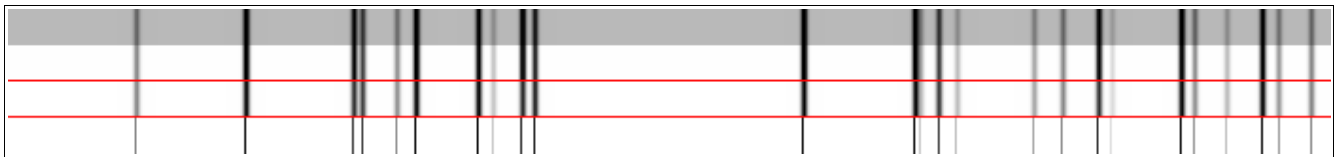
En la tabla 4.12 se muestran los parámetros  $\mu$  que describen al carril  $S_c$ . De igual forma se muestran los parámetros de la estimación  $E_c$ , la cual contiene una cantidad de bandas igual a la contenida en el carril  $S_c$ .

**Tabla 4.12:** Ubicación de bandas real y estimada para la medición de la desviación promedio de las bandas

Banda	$\mu$ (píxeles)		$ S_c - E_c $
	$S_c$	$E_c$	
1	67,6473	67,64785852282	0,0005
2	125,646	125,6625537665	0,0165
3	182,739	182,7409728758	0,002
4	187,274	187,2702830142	0,0037
5	205,532	205,5306700029	0,0015
6	215,728	215,7142432794	0,0137
7	248,608	248,6067449616	0,0012
8	256,459	256,4493157084	0,0096
9	271,955	271,9634716859	0,0085
10	278,473	278,4769623450	0,0039
11	420,798	420,7861601729	0,0118
12	479,628	479,8192193763	0,1912
13	482,44	482,7400695630	0,3001
14	492,197	492,1977915196	0,0008
15	501,954	501,9635734031	0,0096
16	542,639	542,6329902006	0,0060
17	557,859	557,8552046994	0,0038
18	576,86	576,8402920150	0,0197
19	583,909	583,9373224390	0,0283
20	620,664	620,6300742856	0,0339
21	627,593	627,5949325958	0,019
22	644,647	644,6366241135	0,0104
23	663,442	663,4429345640	0,0009
24	671,996	672,0000113064	0,004
25	689,243	689,2471432225	0,0041
-	-	$\Sigma$	0,7051



Para el cálculo de la desviación promedio entre las bandas estimadas y las reales, cada banda del carril  $S_c$  es asociada a la banda dentro del carril  $E_c$  que mejor estime su parámetro  $\mu$ , es decir, la banda en  $E_c$  que está menos alejada de la banda en  $S_c$ . Debido a la exactitud con la que se estimaron las bandas esta asociación da como resultado que cada banda presente en  $S_c$  esta asociada con su análoga en  $E_c$ . Se calcula para cada par de bandas correspondientes la distancia existente entre ellas y se calcula la sumatoria de las distancias de cada par de bandas, para la cual se obtuvo un valor de 0.71 píxeles, que significa una desviación promedio de la ubicación de la banda de 0,029 píxeles. En la figura 4.18 se ilustra el carril real, el carril real sin fondo, el carril estimado por el algoritmo y la ubicación central de cada banda estimada.



**Figura 4.18:** Carriles involucrados en la estimación promedio de las bandas. El carril superior corresponde al carril original, seguido del mismo sin fondo, la estimación y la ubicación central de las bandas estimadas

## 4.8 Rendimiento

La evaluación del rendimiento del algoritmo se basa en tres criterios: el tiempo necesario para encontrar las bandas dentro de un carril, la cantidad de veces evaluada la función objetivo y el número de bandas encontradas por el algoritmo. Los resultados se obtienen del análisis de 10 carriles sintéticos de tamaño  $700 \times 20$ , con 30 bandas a lo largo de cada carril, todas con  $\sigma=1,13$  píxeles. Cada uno de estos es evaluado con dos tamaños diferentes de ventana:  $w=60$  y  $w=120$  píxeles. De forma similar se evalúa el caso en que la función objetivo considera 4 y 8 bandas. Los resultados promedio obtenidos para las posibles combinaciones son mostrados en la tabla 4.13. Las pruebas se realizan utilizando ocho procesadores.

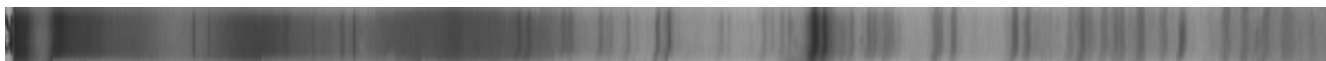
**Tabla 4.13:** Iteraciones y tiempo promedio necesarios para la detección automática de bandas en carriles sintéticos

$w$ (píxeles)	4 bandas			8 bandas		
	Iteraciones	Tiempo (s)	Número de Bandas encontradas	Iteraciones	Tiempo (s)	Número de Bandas encontradas
60	3417979	11,16	28,7	8060245	70,02	29,7
120	1581453	7,5	22,2	4277803	51,18	26,9

Los resultados con menor error de detección de bandas se obtienen para las ventanas de tamaño  $w=60$  píxeles, ya que para el caso en el que la función objetivo considera 8 bandas por ventana se logra un error de 1%, mientras que para el caso en el que se consideran 4 bandas por ventana se obtiene un error de 4,33%. Sin embargo, este último análisis requiere un 16% del tiempo necesario por el análisis con 8 bandas ya que la función objetivo es evaluada una menor cantidad de veces.

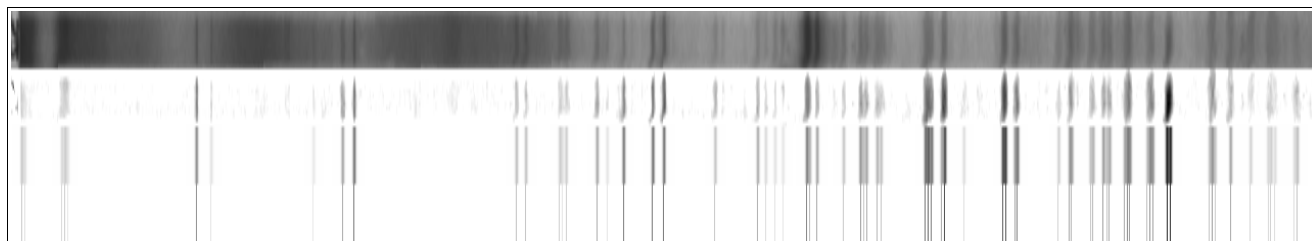
## 4.9 Análisis de carriles completos de geles de electroforesis

Por último se evalúa el método con dos carriles de geles de electroforesis, cada uno con una densidad de bandas diferente, el primero de ellos se ilustra en la figura 4.19. Este carril es el que presenta mayor densidad de bandas, permitiendo observar el comportamiento del algoritmo en situaciones en las cuales existe aglomeración de bandas.



**Figura 4.19:** Carril extraído de un gel de electroforesis para evaluar el comportamiento del algoritmo ante aglomeración de bandas

La estimación se realiza utilizando ventanas de 60 píxeles, un  $\sigma$  de 1,13 píxeles y un valor de umbral de 20ui, considerando en la función objetivo 4 bandas por ventana, se obtiene como resultado los carriles ilustrados en la figura 4.20, los cuales han sido desplegados juntos para facilitar la interpretación de lo realizado por el método propuesto. En orden descendente estos corresponden inicialmente al carril original, el carril original sin el fondo, el carril estimado por método y por último la ubicación central de las bandas detectadas. Para este caso el algoritmo evalúa 5785905 veces la función objetivo en un lapso de tiempo de 30s.



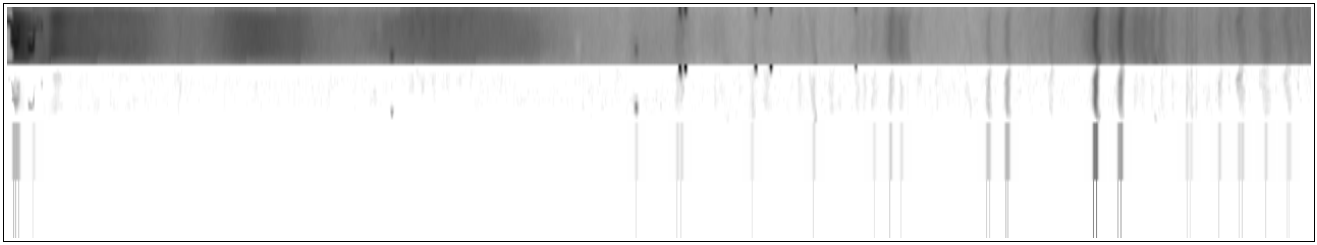
**Figura 4.20:** Carriles involucrados en el proceso de optimización de la función objetivo para el caso de aglomeración de bandas

Por otro lado para evaluar el comportamiento del algoritmo en carriles en los cuales las regiones libres de bandas son mayores que las regiones con bandas se utiliza el carril de la figura 4.21



**Figura 4.21:** Carril extraído de un gel de electroforesis para evaluar el comportamiento del algoritmo si hay mayor cantidad de regiones libres de bandas

obteniendo como resultado de forma similar al caso anterior el grupo de carriles mostrado en la figura 4.22. Evaluando la función objetivo 3608133 veces en un lapso de 20s.



**Figura 4.22:** Carriles involucrados en el proceso de optimización de la función objetivo para el caso en el cual las regiones sin bandas son mayores que las regiones con bandas

Se observa que en ambos casos analizados el sistema propuesto para la detección automática de bandas estima satisfactoriamente la ubicación de las bandas presentes en los carriles.

# Capítulo 5

## 5 Conclusiones y Recomendaciones

### 5.1 Conclusiones

En el presente trabajo se proporciona una estrategia para realizar la ubicación automática de bandas de los geles de electroforesis. Se comprobó que es posible modelar la distribución de intensidad que caracteriza el perfil de una banda mediante una función gaussiana, para la cual su valor medio determina la ubicación central de la banda a lo largo del carril. Esto permitió modelar la distribución de bandas como una sumatoria de funciones gaussianas para así encontrar la ubicación de las mismas mediante una estrategia de optimización de parámetros y una función objetivo. Para encontrar la ubicación de las bandas es necesario realizar la extracción del fondo de la imagen.

Por otra parte, es posible encontrar los parámetros de las bandas estableciendo como función objetivo el error cuadrático medio entre la sumatoria de funciones gaussianas y la intensidad promedio del carril. Se comprobó que para encontrar los parámetros que minimizan esta función objetivo es necesario utilizar un optimizador basado únicamente en la información dada por la función y una estrategia de generación de puntos multidimensionales iniciales a optimizar, lo cual es posible combinando los principios de los algoritmos genéticos PESA y los frentes de Pareto con una estrategia de optimización como Downhill Simplex.

Para realizar una estimación acertada de la cantidad de bandas presentes en el carril y evitar el duplicado de bandas es necesario un algoritmo para la fusión de bandas.

Finalmente la estrategia de detección automática de bandas presentada en este trabajo permite obtener la ubicación central de las bandas con una precisión decimal según el tipo de variable utilizada para la implementación del algoritmo.

## 5.2 Recomendaciones

- La exactitud con la que el presente trabajo realiza la ubicación de las bandas presentes en el carril depende del grado de confiabilidad con el cual se determina el  $\sigma$  de las bandas. Se recomienda implementar un algoritmo de estimación de  $\sigma$  más robusto para mejorar los resultados obtenidos durante la detección de las bandas.
- Se recomienda para trabajos posteriores la implementación de un optimizador más eficiente que Downhill Simplex en cuanto a cantidad de evaluaciones de la función se refiere.

# Bibliografía

- [1] X. Ye, C.Y. Suen, M. Cheriet y E. Wang. 1999.A Recent Development in Image Analysis of Electrophoresis Gels. Vision Interface '99.Canada.
- [2] Michael Lehmann. 2007.Agarose Gel Electrophoresis. Freshmen Biology Section.University of Arkansas.
- [3] P. Alvarado, A. Salazar, O. Murillo, F. Rojas, J. Peraza. 2009.Análisis por computador de imágenes de geles de electroforesis: métodos avanzados de manejo de meta-información y procesamiento digital de imágenes. Propuesta de Proyecto de Investigación.ITCR.
- [4] P. Alvarado, A. Salazar, O. Murillo, F. Rojas, J. Peraza. 2010.Análisis por computador de imágenes de geles de electrofresis para la caracterización molecular de organismos. Informe Final.Vicerrectoría de Investigación y Externsión, ITCR.
- [5] R. García . 2009.Corrección de distorsión geométrica y detección de carriles en imágenes degeles de electroforesis para la caracterización molecular de organismos por computador.. Informe de Proyecto de Graduación.ITCR.
- [6] B. Lomonte. 2007.Manual de Métodos Inmunológicos. Capítulo 13: Electroforesis en Gel de Poliacrilamida. Instituto Clodomiro Picado, Facultad de Microbiología, UCR.
- [7] A. Aguilar. 2010.Detección y corrección del Efecto Sonrisa en imágenes de geles de electroforesis utilizando modelos activos de forma acoplados. Informe de Proyecto de Graduación.ITCR.
- [8] J. Han, C. Moraga. 1995.The influence of the sigmoid function parameters on the speed of backpropagation learning. Computational Models of Neurons and Neural Nets.Research Group Computational Intelligence Dept. of Computer Science, University of Dortmund.
- [9] M.T. Tommiska. 2003.Efficient digital implementation of the sigmoid function for reprogrammable logic. Proceedings Computers and Digital Techniques.IEE.

- [10] M. Vílchez. 2009. Variables aleatorias discretas. Capítulo 4. ITCR.
- [11] M. Isabel Ribeiro. 2004. Gaussian Probability Density Functions: Properties and Error Characterization. Technical Report. Institute for Systems and Robotics.
- [12] John A. Luckey, Tracy B. Norris, Lloyd M. Smith. 1993. Analysis of resolution in DNA sequencing by capillary gel electrophoresis. 3067-3075 .The Journal of Physical Chemistry .
- [13] Rongsheng Lina, David T. Burkeb, Mark A. Burns. 2003. Selective extraction of size-fractionated DNA samples in microfabricated electrophoresis devices. Journal of Chromatography A. University of Michigan.
- [14] Z. Wang and A. C. Bovik. 2009. Mean Squared Error: Love it or Leave it?—A New Look at Fidelity Measures. IEEE Signal Process. Mag., vol. 26. IEEE.
- [15] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery. 2007. Numerical Recipes 3rd Edition. The Art of Scientific Computing. Cambridge University Press.
- [16] Christopher M. Bishop. 1995. Neural Networks for Pattern Recognition. Institute for Adaptive and Neural Computation. Oxford University Press.
- [17] Nelder, J.A., Mead, R.. 1965. A Simplex Method for Function Minimization. vol. 7, pp. 308–313. Computer Journal.
- [18] Charles H.-T. Wang . 2006. Mathematical Optimization . EG5090. University of Aberdeen.
- [19] M. R. Everingham, H. Muller, B. T. Thomas. 2002. Evaluating image segmentation algorithms using the Pareto Front. volume IV of LNCS 2353, pages 34–48. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, Proceedings of the 7th European Conference on Computer Vision.
- [20] Alvarado, Pablo. 2004. Segmentation of color images for interactive 3D object retrieval. Tesis doctoral. RWTH-Aachen.
- [21] D. W. Corne, J. D. Knowles, M. J. Oates. 2000. The Pareto envelope-based selection algorithm for multiobjective optimization. In M. S. et al. (Ed.), Parallel Problem Solving from Nature. PPSN VI, Berlin, pp. 839–848.
- [22] Bryan S. Morse. 2000. Thresholding. Lecture 4. Brigham Young University.
- [23] Carballido, N., García, A., Ballarín, V., Pastore. 2005. Desarrollo de técnicas de procesamiento digital para la optimización de imágenes de fragmentos de ADN obtenidas en estudios de identificación humana. Facultad de Ingeniería. Universidad Nacional de Mar de Plata.
- [24] Bajla, I., Holländer, I. Burg, K. 2001. Improvement of Electrophoretic Gel Image Analysis. Measurement Science Review. Slovak Academy of Science.

- [25] Glasbey C, Vali L, Gustafsson J. 2005. A statistical model for unwarping of 1-D electrophoresis gels. *Electrophoresis*. Nov;26(22):4237-42.
- [26] Li, L. and Speed, T. P. 2000. Parametric deconvolution of positive spike trains. 1279-1301. *The Annals of Statistics*.
- [27] Gonzalez, Rafael C. and Woods, Richard E. 2006. *Digital Image Processing*. 3rd Edition. Prentice-Hall, Inc.
- [28] G.Q. Yin and D. Bruckner. 2009. Gaussian Mixture Models and Split-Merge Algorithm for Parameter Analysis of Tracked Video Objects. 35th IEEE IECON'09. Vienna University of Technology.